TRABAJO DE FIN DE MÁSTER

Máster Interuniversitario en Estadística e Investigación Operativa

Título: Análisis de regresión beta a través de distancias

Autor: Oscar Orlando Melo Martínez

Directores: Josep Maria Oller Sala
Francesc Oliva Cuyas

Departamento: Departamento de Estadística e Investigación Operativa

Convocatoria: **10 / junio / 2010**

Facultat de Matemàtiques i Estadística
UNIVERSITAT POLITÈCNICA DE CATALUNYA

UNIVERSITAT DE BARCELONA

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Tesis de master

Análisis de regresión beta a través de distancias

Oscar Orlando Melo Martínez

Director: Josep Maria Oller Sala y Francesc Oliva Cuyas

Departamento de Estadística e Investigación Operativa

A mis padres: María Martínez y Gustavo Melo

Resumen

En este trabajo, se propone una metodología basada en distancias con la finalidad de ajustar variables respuesta tipo beta de precisión constante y covariables de dispersión; se presenta el modelo propuesto, se realiza la estimación de los diferentes parámetros involucrados por el método de máxima verosimilitud, se hace inferencia para muestras grandes y se lleva a cabo el proceso de validación del modelo propuesto. A partir del enfoque planteado se hace una aplicación en donde se ajustan los modelos de regresión beta de distancias con precisión constante y variable a partir del uso de la distancia de Gower porque las variables explicativas son mixtas.

Palabras clave: distancia Gower, modelo beta, estimación máximo verosímil, modelos de precisión constante y variable

MSC2000: Codis de la *Mathematic Subject Classification*

Abstract

In this work, the beta regression with constant precision parameter and dispersion covariates is mixture with the methodology based on distances in order to adjust beta outcome variables: I let the regression structure to be nonlinear, and I allow a regression structure for the precision parameter, which may also be nonlinear. In addition, the proposed model is presented and estimated. The different parameters involved are estimated by the method of maximum likelihood, the inference is made for large samples, and the validation of the proposed model is carried out. Finally, an application, where the beta regression model through Gower distances is used, is presented.

Keywords: Gower distance, beta model, maximum likelihood estimation, models of constant and variable precision

MSC2000: 2000Mathematical Subject Classification

Índice general

Introducción

No son pocos los fenómenos de la vida cotidiana que se salen de las manos al intentar traducirlos a un lenguaje simbólico propio de la disciplina estadística. En consecuencia, se cae en el abismo de ajustar dichos fenómenos a los modelos que se tiene a disposición, en lugar de permitir que los datos "hablen por sí solos". El intento por acercar la teoría a las situaciones reales ha motivado el desarrollo de técnicas estadísticas encaminadas a encontrar modelos cada vez más generales que respondan fielmente a los objetivos del investigador en correspondencia con la realidad.

En particular cuando el interés radica en relacionar variables explicativas con una respuesta, la primera referencia es el modelo de regresión lineal $E(y) = X\beta$, quizá una de las herramientas más popular y antigua en estadística; éste se presenta como instrumento útil cuando la normalidad e independencia sobre los errores es innegable, allí se determina el efecto de las covariables a través de las componentes del vector β.

El modelo de regresión lineal no es apropiado en situaciones en las cuales la respuesta está restringida al intervalo (0,1), debido a que el método de estimación de mínimos cuadrados ordinarios (MCO) puede generar valores ajustados que excedan dichas cotas inferior y superior. En este caso, como se muestra en Kieschnick & McCullough (2003), Ferrari & Cribari-Neto (2004) y Vasconcellos & Cribari-Neto (2005), una posible solución es transformar la variable dependiente para asumir que ésta toma valores sobre la recta real, y luego, modelar la media de la respuesta transformada como un predictor lineal basado en un conjunto de variables exógenas.

La distribución beta es muy flexible para el modelamiento de esta clase de datos ya que su función de densidad presenta diferentes formas en función de los valores de los parámetros. En particular, Bury (1999) enumera las aplicaciones de la distribución beta en la ingeniería y Johnson, Kotz & Balakrishnan (1995) presentan y debaten una serie de aplicaciones de la distribución beta. De acuerdo a los anteriores autores, esta distribución se encuentra entre las más frecuentemente empleadas en el modelamiento de distribuciones teóricas. Así mismo, Krysicki (1999) presenta algunas nuevas propiedades de esta distribución.

El modelo de regresión propuesto en Ferrari & Cribari-Neto (2004) es generado para situaciones donde la variable respuesta y es continua y definida sobre el intervalo unitario estandarizado $0 < y < 1$ y, la estructura de regresión involucra

regresores y parámetros desconocidos. Ospina, Cribari-Neto & Vasconcellos (2006) obtienen el sesgo de segundo orden de los estimadores de maxima verosimilitud y los utilizan para definir los estimadores de sesgo ajustado, los cuales son muy útiles para solucionar el problema en muestras pequeñas.

En el *modelo de regresión beta* propuesto por Ferrari & Cribari-Neto (2004) los parámetros de regresión son interpretables en términos de la media de y (la variable de interés) y el modelo es, naturalmente, heterocedastico y es acomodado fácilmente a las asimetrías. Una variante del modelo de regresión beta que permite modelar la no linealidad y la variable de dispersión fue propuesta por Simas, Barreto-Souza & Rocha (2010). En particular, en este modelo más general, el parámetro que representa la precisión de los datos no se supone que es constante a través de las observaciones, sino que puede variar, esto se conoce como *modelo de regresión beta de precisión variable*.

Por otro lado, muchos métodos de estadística y análisis de datos utilizan el concepto geométrico de distancia entre individuos o poblaciones, estos métodos se aplican en campos tales como la antropología, biología, genética, psicología, entre otros (Arenas & Cuadras 2002). Las distancias, aparecen en muchos aspectos de la estadística: contraste de hipótesis, estimación, regresión, análisis discriminante, etc. (Cuadras 2007). Cuadras & Arenas (1990) proponen el método de regresión múltiple basado en el análisis de distancias utilizando diferentes métricas para el trabajo con variables explicativas continuas y categóricas. Cuadras, Arenas & Fortiana (1996) presentaron algunos resultados adicionales del modelo basado en distancias (DB) para la predicción de variables mezcladas (continuas y categóricas) y exploran el problema de información faltante dando una solución utilizando DB. Uno de los trabajos más recientes es el de Esteve, Boj & Fortiana (2010), quienes proponen un método donde incluyen términos polinomiales y de interacción en la regresión basada en distancias, bajo las propiedades de un producto de matrices semi-Hadamard o Khatri-Rao.

En este trabajo, se enlaza los modelos de regresión beta con el método de distancias, en donde todas las variables explicativas son linealmente independientes. Se discute la estimación de los parámetros desconocidos por el método de máxima verosimilitud, algunas técnicas de diagnóstico y se considera la inferencia para muestras grandes. El modelamiento y los procedimientos inferenciales propuestos son similares a los modelos lineales generalizados discutidos por McCullagh & Nelder (1989) y Myers, Montgomery & Vinning (2002), excepto que la distribución de la respuesta no es miembro de la familia exponencial. Aunque la respuesta no es miembro de la familia exponencial se hace una adaptación a esta familia siguiendo las propuestas realizadas en modelos generalizados por McCulloch & Searle (2001), Lee & Nelder (2002), Dobson (2002) y Smith & Ridout (2003).

Este trabajo está desarrollado en tres capítulos, en el primero se plantea el modelo de regresión beta de distancias con precisión constante y variable, además se hace la estimación de los parámetros involucrados en el modelo propuesto. En el segundo capítulo se realiza el proceso de inferencia del modelo propuesto, al igual que la validación de algunos supuestos deseables como en el modelo lineal de regresión tradicional. En el tercero, se muestra una aplicación de la metodología planteada, concretamente se considera los datos de petróleo convertido a gasolina recolectados

por Prater (1956). Finalmente, se presentan algunas conclusiones y recomendaciones de este trabajo.

Capítulo 1
Planteamiento del modelo y estimación

En este capítulo se hace el planteamiento del modelo de regresión beta de distancias con dispersión constante y variable. Para este fin, se construye el modelo beta con distancias, se estiman los parámetros del modelo y se presenta el modelo de precisión constante como caso especial del modelo de recisión variable.

1.1. Introducción

El modelo se basa en el supuesto que la respuesta tiene distribución beta, la función de densidad beta está dada por

$$\pi(y,p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{(p-1)}(1-y)^{(q-1)}, \qquad 0<y<1$$

donde $p>0$, $q>0$ y $\Gamma(.)$ es la función Gamma. La media y la varianza de y son, respectivamente,

$$E(y) = \frac{p}{p+q} \tag{1}$$

y

$$Var(y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{2}$$

Las estimaciones de p y q por máxima verosimilitud y la aplicación de los ajustes en los sesgos en pequeñas muestras de los estimadores de máxima verosimilitud para los parámetros son analizados en Cribari-Neto & Vasconcellos (2002).

Con el fin de obtener una estructura de regresión para la media de la respuesta junto con el parámetro de precisión, se trabaja con una parametrización diferente de la densidad beta. De acuerdo a Ferrari & Cribari-Neto (2004), sea $\mu = p/(p+q)$ y $\phi = p+q$, es decir, $p = \mu\phi$ y $q = (1-\mu)\phi$. De lo anterior para las ecuaciones (1) y (2) se sigue que $E(y) = \mu$ y $var(y) = V(\mu)/(1+\phi)$, donde $V(\mu) = \mu(1-\mu)$ es la función de varianza. μ es la media de la variable respuesta y ϕ puede ser interpretado como un parámetro de precisión en el sentido de que para un μ fijo, un valor grande de ϕ conlleva a un menor valor de la varianza de y.

En términos de la nueva parametrización, la densidad de y se puede reescribir como

$$f(y,\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{(\mu\phi-1)}(1-y)^{((1-\mu)\phi-1)}, \qquad 0<y<1 \tag{3}$$

donde $0<\mu<1$ y $\phi>0$. Las densidades de la distribución pueden tener diferentes formas dependiendo de los valores de estos dos parámetros. En particular, ésta puede ser simétrica cuando $\mu=1/2$ o asimétrica cuando $\mu\neq 1/2$. Adicionalmente, la dispersión de la distribución para un μ fijo decrece a medida que ϕ crece. En particular cuando $\mu=1/2$ y $\phi=2$ la densidad se reduce a la *distribución uniforme*.

Aunque en este artículo la respuesta está restringida al intervalo unitario $(0,1)$, el modelo propuesto es útil para situaciones donde la respuesta está restringida al intervalo (a,b) donde a y b son escalares conocidos, con $a<b$. En este caso se debe modelar $(y-a)/(b-a)$ en cambio de modelar y directamente. Además, si y también asume los extremos 0 y 1, una transformación útil en la práctica es $(y(n-1)+0{,}5)/n$, donde n es el tamaño de la muestra Smithson & Verkuilen (2006).

Sean $y_1,y_2,\ldots,y_n$ variables aleatorias independientes, donde cada y_i sigue la densidad mostrada en (3) con media μ_i y precisión desconocida ϕ, $i=1,\ldots,n$. El modelo se obtiene asumiendo que la media de y_i puede ser escrita como

$$\eta_i=g(\mu_i)=v_i^t\zeta=\sum_{j=0}^{p}v_{ij}\zeta_j \tag{4}$$

donde $\zeta^t=(\zeta_0,\zeta_1,\ldots,\zeta_p)$ es un vector de $p+1$ parámetros de regresión desconocidos $\zeta\in\Re^{p+1}$, $v_i^t=(v_{i0},v_{i1},\ldots,v_{ip})$ es un vector de observaciones de $p+1$ covariables ($p+1<n$, con $v_{i0}=1$), las cuales se asumen fijas y conocidas. Finalmente, $g(\cdot)$ es una función de enlace estrictamente monótona y doblemente diferenciable, en donde a cada elemento en el intervalo $(0,1)$ le asigna un número en los reales $\Re$. Existen muchas escogencias posibles para la función de enlace $g(\cdot)$ en el modelo presentado en (4). Por ejemplo, es posible utilizar las funciones: logit, $g(\mu)=\log\{\mu/(1-\mu)\}$; probit, $g(\mu)=\Phi^{-1}(\mu)$ donde $\Phi(\cdot)$ es la función de distribución acumulada de una variable aleatoria normal estándar; complemento log-log, $g(\mu)=\log\{-\log(1-\mu)\}$; y log-log, $g(\mu)=-\log\{-\log(\mu)\}$. Una rica discusión de las funciones de enlace son presentadas en Atkinson (1985) y McCullagh & Nelder (1989).

1.2. Construcción del modelo beta con distancias

En forma matricial el modelo (4) se puede reescribir como:

$$\eta=g(\mu)=V\xi \tag{5}$$

donde $V=(V_1 \quad V_2)$ con V_1 una submatriz de variables continuas y V_2 una submatriz de variables cualitativas. De acuerdo a Cuadras & Arenas (1990) se puede definir la similaridad como:

$$s_{ij}=\frac{\sum_{h=1}^{p_1}\left(\frac{1-|v_{ih}-v_{jh}|}{G_h}\right)+a+\alpha}{p_1+(p_2-d)+p_3} \tag{6}$$

donde p_1 es el número de variables continuas, a y d son el números de coincidencias y no coincidencias para las p_2 variables binarias, respectivamente, y α es el número

de coincidencias de las p_3 variables cualitativas. G_h es el rango de la h-ésima variable cualitativa. La semejanza (6) es conocida como distancia de Gower (1968).

La distancia al cuadrado entre los individuos i y j es:

$$d_{ij}^2 = 1 - s_{ij}$$

Ahora haciendo $D = (d_{ij})$ una matriz de distancias Euclideana sobre el conjunto de n individuos.

Si todas las variables explicativas en (5) son continuas sería mejor definir la distancia:

$$d_{ij}^2 = (v_i - v_j)^t (v_i - v_j)$$

o de un modo bastante eficiente, la distancia valor absoluto:

$$\delta_{ij}^2 = \sum_{h=1}^{p} |v_{ih} - v_{jh}|$$

En el caso que todas las variables explicativas en el modelo (5) sean cualitativas una medida bastante utilizada de similaridad entre dos individuos i y j es m_{ij}, el número de estados presentes simultáneamente en i y en j. Como $m_{ij} \leq p$, una medida de distancia viene dada por

$$\delta_{ij}^2 = 2(p - m_{ij})$$

Una vez seleccionada alguna de las distancias presentadas anteriormente se define $A = (a_{ij})$ donde $a_{ij} = -d_{ij}^2/2$ y $B = HAH$ con $H = I - \frac{1}{n}11^t$ y 1 un vector de unos de tamaño $n \times 1$. Se sabe que B es una matriz semi-definida positiva (Mardia, Kent & Bibby 1979) de rango p, de este modo

$$\begin{aligned} B &= \left(I - \frac{1}{n}11^t\right) A \left(I - \frac{1}{n}11^t\right) \\ &= XX^t = \Lambda = diag(\lambda_1, \ldots, \lambda_p) \end{aligned}$$

donde X es una matriz de $n \times p$ de rango p y los λ_i's (i=1,...,p) son los valores propios positivos de B.

Además las filas $x_1^t, \ldots, x_n^t$ de la matriz X son las coordenadas principales de B con respecto a la distancia D y como un individuo i es parecido a un individuo j en (5) entonces $v_i \cong v_j$, y por lo tanto $x_i \cong x_j$.

Particionando $X = (X_{(k)} \quad L)$ donde $X_{(k)}$ contiene un suconjunto de k columnas de X. De esta manera el modelo (5) se puede reescribir como

$$\eta = \beta_0 1 + X_{(k)}\beta_{(k)} \tag{7}$$

donde $X_{(k)} = (X_1, \ldots, X_k)$ con cada X_i, $i = 1, \ldots, p$ una columna de X, siendo cada X_i una componente principal.

El modelo propuesto en (7) se puede escribir de la siguiente forma

$$\eta = \beta_0 1 + \sum_{j=1}^{k} \beta_j X_j \tag{8}$$

y expresando individuo a individuo es

$$\begin{aligned}\eta_i =& g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \sum_{j=0}^{k} x_{ij}\beta_j \\ =& x_i^t \beta\end{aligned}$$

con $x_{i0} = 1$, $x_i^t = (x_{i0}, \ldots, x_{ik})$ y $\beta^t = (\beta_0, \ldots, \beta_k)$.

1.3. Estimación de parámetros

Retomando los resultados de la regresión beta presentados en Ferrari & Cribari-Neto (2004), se considera la densidad para cada y_i dada en (3), el logaritmo de la función de verosimilitud es:

$$\begin{aligned}l_i(\mu_i, \phi) =& \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) \\ &+ (\mu_i\phi - 1)\log(y_i) + ((1-\mu_i)\phi - 1)\log(1-y_i)\end{aligned}$$

Una extensión de la anterior propuesta empleada por Smithson & Verkuilen (2006) y formalmente introducida (junto con otras extensiones) por Simas et al. (2010) es la del modelo de regresión beta con dispersión variable. En este modelo, el parámetro de precisión no es constante para todas las observaciones, sino que en cambio es modelado de forma similar al parámetro de la media. Más específicamente, $y_i \sim \mathfrak{B}(\mu_i, \phi_i)$ variables aleatorias independientes, $i = 1, \ldots, n$, y

$$g_1(\mu_i) = \eta_{1i} = f_1(x_i^t, \beta) \qquad g_1(\phi_i) = \eta_{2i} = f_2(z_i^t, \alpha) \tag{9}$$

donde $\beta = (\beta_1, \ldots, \beta_k)^t$, $\alpha = (\alpha_1, \ldots, \alpha_h)^t$ son vectores correspondientes al conjunto de parámetros desconocidos y que se asumen funcionalmente independientes, $\beta \in \Re^k$ y $\alpha \in \Re^h$,$k+h < n$, η_{1i} y η_{2i} son los predictores lineales, y, $x_i^t = (x_{i1}, ..., x_{iq_1})$ y $z_i^t = (z_{i1}, ..., z_{iq_2})$ son los vectores de observaciones de q_1 y q_2 covariables conocidas, respectivamente.

Se asumirá que las matrices de derivadas $X = \partial\eta_1/\partial\beta$ y $Z = \partial\eta_2/\partial\alpha$ tienen rango k y h, respectivamente, y que las funciones de enlace $g_1 : (0,1) \to \Re$ y $g_2 : (0,\infty) \to \Re$ son estrictamente monótonas y doblemente diferenciables. Algunas de las funciones de enlace para $g_1(\mu)$ se presentaron anteriormente, y para g_2 se tienen: el logaritmo, $g_2(\phi) = \log\phi$; la función raíz cuadrada, $g_2(\phi) = \sqrt{\phi}$; y la función identidad, $g_2(\phi) = \phi$, entre otras.

La función de log-verosimilitud para esta clase de modelos de regresión beta tienen la forma:

$$l(\beta, \alpha) = \sum_{i=1}^{n} l_i(\mu_i, \phi_i) \tag{10}$$

donde

$$\begin{aligned}l_i(\mu_i, \phi_i) =& \log\Gamma(\phi_i) - \log\Gamma(\mu_i\phi_i) - \log\Gamma((1-\mu_i)\phi_i) \\ &+ (\mu_i\phi_i - 1)\log(y_i) + ((1-\mu_i)\phi_i - 1)\log(1-y_i)\end{aligned} \tag{11}$$

con $\mu_i = g_1^{-1}(\eta_{1i})$ y $\phi_i = g_2^{-1}(\eta_{2i})$ definidos como se indica en (9). Derivando parcialmente con respecto a cada uno de los β_j para $j = 1, \ldots, k$, se obtiene:

Retomando la ecuación (10) y derivando parcialmente con respecto a cada uno de los β_j, para $j = 1, \ldots, p$, se obtiene

$$U_j(\beta, \alpha) = \frac{\partial l(\beta, \alpha)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i, \phi_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_{1i}} \frac{\partial \eta_{1i}}{\partial \beta_j} \tag{12}$$

donde $\partial \mu_i / \partial \beta_j = x_{ij}$ y de la ecuación (9), $\partial \mu_i / \partial \eta_{1i} = 1/g_1'(\mu_i) = dg_1^{-1}(\eta_{1i})/d\eta_{1i}$. Adicionalmente de la ecuación (11) se tiene que

$$\begin{aligned} \frac{\partial l_i(\mu_i, \phi_i)}{\partial \mu_i} =& -\phi_i \frac{\partial \log \Gamma(\mu_i \phi_i)}{\partial \mu_i} + \phi_i \frac{\partial \log \Gamma((1-\mu_i)\phi_i)}{\partial \mu_i} + \phi_i \log(y_i) - \phi_i \log(1-y_i) \\ =& \phi_i \left(\log \frac{y_i}{1-y_i} - (\psi(\mu_i \phi_i) - \psi((1-\mu_i)\phi_i)) \right) \end{aligned} \tag{13}$$

donde $\psi(\cdot)$ es la función digamma, es decir, $\psi(z) = \partial \log \Gamma(z)/\partial(z)$ para $z > 0$.

Sean $y_i^* = \log(y_i/(1-y_i))$ y $\mu_i^* = \psi(\mu_i \phi_i) - \psi((1-\mu_i)\phi_i)$, entonces

$$U_j(\beta, \alpha) = \frac{\partial l(\beta, \alpha)}{\partial \beta_j} = \sum_{i=1}^{n} \phi_i (y_i^* - \mu_i^*) \frac{1}{g_1'(\mu_i)} x_{ij}$$

Ahora, derivando parcialmente con respecto a cada uno de los α_s, para $s = 1, \ldots, h$, en la ecuación (10), se obtiene:

$$U_s(\beta, \alpha) = \frac{\partial l(\beta, \alpha)}{\partial \alpha_s} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i, \phi_i)}{\partial \phi_i} \frac{\partial \phi_i}{\partial \eta_{2i}} \frac{\partial \eta_{2i}}{\partial \alpha_s} \tag{14}$$

donde $\partial \phi_i / \partial \alpha_s = z_{is}$ y de la ecuación (9), $\partial \phi_i / \partial \eta_{2i} = 1/g_2'(\phi_i) = dg_2^{-1}(\eta_{2i})/d\eta_{2i}$. Adicionalmente de la ecuación (11) se tiene que

$$\begin{aligned} \frac{\partial l_i(\mu_i, \phi_i)}{\partial \phi_i} =& \frac{\partial \log \Gamma(\phi_i)}{\partial \phi_i} - \mu_i \frac{\partial \log \Gamma(\mu_i \phi_i)}{\partial \phi_i} - (1-\mu_i) \frac{\partial \log \Gamma((1-\mu_i)\phi_i)}{\partial \phi_i} \\ & + \mu_i \log(y_i) + (1-\mu_i) \log(1-y_i) \\ =& \psi(\phi_i) - \mu_i(\psi(\mu_i \phi_i) - \psi((1-\mu_i)\phi_i)) - \psi((1-\mu_i)\phi_i) \\ & + \mu_i \log \frac{y_i}{1-y_i} + \log(1-y_i) \\ =& \psi(\phi_i) + \mu_i(y_i^* - \mu_i^*) - \psi((1-\mu_i)\phi_i) + \log(1-y_i) \end{aligned} \tag{15}$$

Por lo tanto,

$$U_s(\beta, \alpha) = \sum_{i=1}^{n} (\psi(\phi_i) + \mu_i(y_i^* - \mu_i^*) - \psi((1-\mu_i)\phi_i) + \log(1-y_i)) \frac{1}{g_2'(\phi_i)} z_{is}$$

Bajo condiciones de regularidad se tiene que:

$$E\left(log \frac{y_i}{1-y_i} \right) = \psi(\mu_i \phi_i) - \psi((1-\mu_i)\phi_i) \quad \text{y} \quad E[log(1-y_i)] = \psi((1-\mu_i)\phi_i) - \psi(\phi_i)$$

Considere el vector completo de parámetros $\theta = (\beta^t, \alpha^t)^t$. Definiendo los vectores $y^* = (y_1^*, \ldots, y_n^*)^t$ y $\mu^* = (\mu_1^*, \ldots, \mu_n^*)^t$, $v = (v_1, \ldots, v_n)^t$, las matrices

$T_1 = diag(d\mu_i/d\eta_{1i})$, $T_2 = diag(d\phi_i/d\eta_{2i})$ y $\Upsilon = diag(\phi_i)$, con $diag(\mu_i)$ denotando la matriz diagonal $n \times n$ con elementos μ_i, $i = 1, ..., n$, y donde $v_i = \mu_i(y_i^* - \mu_i^*) + \psi(\phi_i) - \psi((1-\mu_i)\phi_i) + log(1-y_i)$. Por lo tanto, la expresión matricial de la función Score obtenida por diferenciación de la función log-verosímil con respecto a los parámetros desconocidos está dada por $U(\theta) = (U_\beta(\beta,\alpha)^t, U_\alpha(\beta,\alpha)^t)^t$, donde

$$U_\beta(\beta,\alpha) = X^t \Upsilon T_1 (y^* - \mu^*) \quad \text{y} \quad U_\alpha(\beta,\alpha) = Z^t T_2 v \tag{16}$$

El siguiente paso es hallar la segunda derivada de $l(\beta,\alpha)$ con respecto a los β's, con la finalidad de obtener una expresión para la matriz de información de Fisher. Derivando (12) con respecto a los β se encuentra:

$$\begin{aligned}\frac{\partial^2 l(\beta,\alpha)}{\partial\beta_j \partial\beta_{j'}} &= \sum_{i=1}^{n} \frac{\partial}{\partial\mu_i}\left(\frac{\partial l_i(\mu_i,\phi_i)}{\partial\mu_i}\frac{d\mu_i}{d\eta_{1i}}\right)\frac{d\mu_i}{d\eta_{1i}}\frac{\partial\eta_{1i}}{\partial\beta_j}x_{ij'} \\ &= \sum_{i=1}^{n}\left(\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\mu_i^2}\frac{d\mu_i}{d\eta_{1i}} + \frac{\partial l_i(\mu_i,\phi_i)}{\partial\mu_i}\frac{\partial}{\partial\mu_i}\frac{d\mu_i}{d\eta_{1i}}\right)\frac{d\mu_i}{d\eta_{1i}}x_{ij}x_{ij'}\end{aligned}$$

Como $E(\partial l_i(\mu_i,\phi_i)/\partial\mu_i) = 0$, se tiene que

$$E\left(\frac{\partial^2 l(\beta,\alpha)}{\partial\beta_j\partial\beta_{j'}}\right) = \sum_{i=1}^{n} E\left(\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\mu_i^2}\right)\left(\frac{d\mu_i}{d\eta_{1i}}\right)^2 x_{ij}x_{ij'}$$

Derivando parcialmente con respecto a μ_i la ecuación (13) se encuentra que

$$\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\mu_i^2} = -\phi_i^2(\psi'(\mu_i\phi_i) + \psi'((1-\mu_i)\phi_i))$$

y entonces

$$E\left(\frac{\partial^2 l(\beta,\alpha)}{\partial\beta_j\partial\beta_{j'}}\right) = -\sum_{i=1}^{n}\phi_i^2 a_i x_{ij}x_{ij'}$$

con

$$a_i = \{\psi'(\mu_i\phi_i) + \psi'((1-\mu_i)\phi_i)\}\frac{1}{g_1'(\mu_i)^2}$$

donde $\psi'(\cdot)$ es la función trigamma.

De la ecuación (12), la segunda derivada de $l(\beta,\alpha)$ con respecto a α_s, se expresa de la siguiente forma:

$$\begin{aligned}\frac{\partial^2 l(\beta,\alpha)}{\partial\alpha_s\partial\beta_j} &= \sum_{i=1}^{n}\frac{\partial}{\partial\phi_i}\left(\frac{\partial l_i(\mu_i,\phi_i)}{\partial\mu_i}\frac{d\mu_i}{d\eta_{1i}}\right)\frac{d\phi_i}{d\eta_{2i}}\frac{\partial\eta_{2i}}{\partial\alpha_s}x_{ij'} \\ &= \sum_{i=1}^{n}\left(\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\phi_i\partial\mu_i}\frac{d\mu_i}{d\eta_{1i}} + \frac{\partial l_i(\mu_i,\phi_i)}{\partial\mu_i}\frac{\partial}{\partial\phi_i}\frac{d\mu_i}{d\eta_{1i}}\right)\frac{d\phi_i}{d\eta_{2i}}z_{is}x_{ij'}\end{aligned}$$

Ahora tomando esperanza a ambos lados de la anterior expresión, se tiene que

$$E\left(\frac{\partial^2 l(\beta,\alpha)}{\partial\alpha_s\partial\beta_j}\right) = \sum_{i=1}^{n} E\left(\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\phi_i\partial\mu_i}\right)\left(\frac{d\mu_i}{d\eta_{1i}}\right)\left(\frac{d\phi_i}{d\eta_{2i}}\right)z_{is}x_{ij'}$$

Derivando parcialmente con respecto a ϕ_i la ecuación (13) se encuentra que

$$\begin{aligned}\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\phi_i\partial\mu_i} =& y_i^* - \mu_i^* - \phi_i\{\mu_i[\psi'(\mu_i\phi_i) + \psi'((1-\mu_i)\phi_i)] - \psi'((1-\mu_i)\phi_i)\}\\ =& y_i^* - \mu_i^* - \phi_i[\mu_i a_i - \psi'((1-\mu_i)\phi_i)]\end{aligned}$$

y entonces

$$E\left(\frac{\partial^2 l(\beta,\alpha)}{\partial\alpha_s\partial\beta_j}\right) = -\sum_{i=1}^{n}\phi_i[\mu_i a_i - \psi'((1-\mu_i)\phi_i)]\left(\frac{d\mu_i}{d\eta_{1i}}\right)\left(\frac{d\phi_i}{d\eta_{2i}}\right)z_{is}x_{ij'}$$

Por último, derivando parcialmente de nuevo con respecto a ϕ_i la ecuación (14) se encuentra que

$$\begin{aligned}\frac{\partial^2 l(\beta,\alpha)}{\partial\alpha_s\partial\alpha_{s'}} =& \sum_{i=1}^{n}\frac{\partial}{\partial\phi_i}\left(\frac{\partial l_i(\mu_i,\phi_i)}{\partial\phi_i}\frac{d\phi_i}{d\eta_{2i}}\right)\frac{d\phi_i}{d\eta_{2i}}\frac{\partial\eta_{2i}}{\partial\alpha_s}z_{is'}\\ =& \sum_{i=1}^{n}\left(\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\phi_i^2}\frac{d\phi_i}{d\eta_{2i}} + \frac{\partial l_i(\mu_i,\phi_i)}{\partial\phi_i}\frac{\partial}{\partial\phi_i}\frac{d\phi_i}{d\eta_{2i}}\right)\frac{d\phi_i}{d\eta_{2i}}z_{is}z_{is'}\end{aligned}$$

Como $E(\partial l_i(\mu_i,\phi_i)/\partial\phi_i) = 0$, al tomar esperanza a ambos lados de la anterior expresión, se tiene que

$$E\left(\frac{\partial^2 l(\beta,\alpha)}{\partial\alpha_s\partial\alpha_{s'}}\right) = \sum_{i=1}^{n}E\left(\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\phi_i^2}\right)\left(\frac{d\phi_i}{d\eta_{2i}}\right)^2 z_{is}z_{is'}$$

Derivando parcialmente con respecto a ϕ_i la ecuación (15) se encuentra que

$$\begin{aligned}\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial\phi_i^2} =& \psi'(\phi_i) - \mu_i[\mu_i\psi'(\mu_i\phi_i) - (1-\mu_i)\psi'((1-\mu_i)\phi_i)] - (1-\mu_i)\psi'((1-\mu_i)\phi_i)\\ =& \psi'(\phi_i) - \mu_i^2\psi'(\mu_i\phi_i) - (1-\mu_i)^2\psi'((1-\mu_i)\phi_i)\end{aligned}$$

y entonces

$$\begin{aligned}E\left(\frac{\partial^2 l(\beta,\alpha)}{\partial\alpha_s\partial\alpha_{s'}}\right) =& -\sum_{i=1}^{n}[(1-\mu_i)^2\psi'((1-\mu_i)\phi_i) + \mu_i^2\psi'(\mu_i\phi_i) - \psi'(\phi_i)]\left(\frac{d\phi_i}{d\eta_{2i}}\right)^2 z_{is}z_{is'}\\ =& -\sum_{i=1}^{n}b_i\left(\frac{d\phi_i}{d\eta_{2i}}\right)^2 z_{is}z_{is'}\end{aligned}$$

donde $b_i = (1-\mu_i)^2\psi'((1-\mu_i)\phi_i) + \mu_i^2\psi'(\mu_i\phi_i) - \psi'(\phi_i)$.

Los estimadores máximo verosímil (EMV) de β y α se obtienen del sistema no lineal $U(\theta) = 0$, con $\theta = (\beta^t, \alpha^t)^t$. En la práctica, los EMV's se obtienen a través de la maximización numérica de (10) utilizando un algoritmo de optimización no lineal (Newton-Raphson, Fisher's scoring, quasi-Newton, etc). Mayores detalles al respecto se pueden ver en Press, Teukolsky, Vetterling & Flannery (1992). Simas et al. (2010) sugieren como valores iniciales de β y α obtener los estimadores de los siguientes modelos de regresión no-lineal normal con covariables de dispersión:

$$g_1(\mu_i) = f_1(x_i,\beta) \quad \text{y} \quad g_2(\sigma_i^{-2}) = f_2(z_i,\alpha)$$

Este procedimiento producirá $\hat{\beta}^{(0)}$ y $\hat{\alpha}^{(0)}$, los cuales sirven como valores iniciales, donde para encontrar estos valores se asume que $Y_i \sim N(\mu_i, \sigma_i^2)$. Es de observar aquí, que el modelo no lineal normal presentado anteriormente en el sentido del modelo no lineal generalizado, ya que se están utilizando las funciones de enlace.

Definiendo P como la matriz de dimensión $2n \times (k+h)$

$$P = \begin{pmatrix} X & 0 \\ 0 & Z \end{pmatrix} \tag{17}$$

Adicionalmente, sea W la matriz de $2n \times 2n$ definida como

$$W = \begin{pmatrix} W_{\beta\beta} & W_{\beta\alpha} \\ W_{\alpha\beta} & W_{\alpha\alpha} \end{pmatrix} \tag{18}$$

con $W_{\alpha\alpha} = diag\left(b_i\left(\frac{d\phi_i}{d\eta_{2i}}\right)\right)$, $W_{\beta\alpha} = diag\left(\phi_i[\mu_i a_i - \psi'((1-\mu_i)\phi_i)]\left(\frac{d\mu_i}{d\eta_{1i}}\right)\left(\frac{d\phi_i}{d\eta_{2i}}\right)\right)$, $W_{\beta\beta} = diag\left(\phi_i^2 a_i\left(\frac{d\mu_i}{d\eta_{1i}}\right)^2\right)$ y $W_{\alpha\beta} = W_{\beta\alpha}^t$.

Ahora con base en (17) y (18) se obtiene la matriz de información de Fisher para el vector de parámetros θ como:

$$K(\theta) = K(\beta, \alpha) = P^t W P = \begin{pmatrix} K_{\beta\beta} & K_{\beta\alpha} \\ K_{\alpha\beta} & K_{\alpha\alpha} \end{pmatrix} \tag{19}$$

En este caso, como $W_{\beta\alpha} \neq 0$ entonces los parámetros β y α no son ortogonales, en comparación a los modelos lineales generalizados (McCullagh & Nelder 1989), en donde se cumple la ortogonalidad. Sin embargo, los EMC's de $\hat{\theta}$ y $K(\hat{\theta})$ son estimadores consistentes de θ y $K(\theta)$, respectivamente, donde $K(\hat{\theta})$ es la matriz de información de Fisher evaluada en $\hat{\theta}$.

Asumiendo que $J(\theta) = \lim_{n\to\infty} K(\theta)/n$ existe y es no singular, es decir

$$K^{-1} = K^{-1}(\theta) = K^{-1}(\beta, \alpha) = \begin{pmatrix} K^{\beta\beta} & K^{\beta\alpha} \\ K^{\alpha\beta} & K^{\alpha\alpha} \end{pmatrix},$$

Además, se obtiene que $\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N_{k+h}(0, J(\theta)^{-1})$, donde $\xrightarrow{d}$ denota convergencia en distribución. De este modo, si β_j el j-ésimo componente de θ, se encuentra que

$$\left(\hat{\beta}_j - \beta_j\right)\left(K_j^{\beta\beta}\right)^{-1/2} \xrightarrow{d} N(0,1)$$

donde $K_j^{\beta\beta}$ es el j-ésimo elemento de la diagonal de $K^{\beta\beta}$. Sea $K_s^{\alpha\alpha}$ el s-ésimo elemento de la diagonal de $K^{\alpha\alpha}$, entonces si $0 < q < 1/2$ y z_γ representa el cuantil γ de la distribución $N(0,1)$, se tiene que para $j = 1, ..., k$ y $s = 1, ..., h$ los limites asintóticos de confianza para β_j y α_s son

$$\hat{\beta}_j \pm z_{1-q/2}\left(K_j^{\beta\beta}\right)^{1/2} \quad \text{y} \quad \hat{\alpha}_s \pm z_{1-q/2}\left(K_s^{\alpha\alpha}\right)^{1/2}$$

respectivamente, los cuales cubren asintóticamente $100(1-q)\,\%$.

1.4. Modelo de regresión beta de distancias con precisión constante

Para el modelo de regresión beta se tiene en (9), $g_1(\mu_i) = g(\mu_i)$, donde $g(\cdot)$ es una función de enlace, $g_2(\phi_i) = \phi_i$, y además, (9) se puede escribir como:

$$g(\mu_i) = \eta_i = x_i^t\beta \quad \text{y} \quad \phi_i = \phi$$

donde $\phi > 0$ es constante, es decir, se tiene que $Z = \mathbf{1}$. Por lo tanto, el vector score definido en (16) es:

$$U_\beta(\beta, \phi) = \phi X^t T(y^* - \mu^*) \quad \text{y} \quad U_\phi(\beta, \phi) = \sum_{i=1}^{n} v_i \tag{20}$$

donde $T = diag(d\mu_i/d\eta_i)$ y y^*, μ^* y v_i fueron definidos en la sección 1.3. Más aún, las matrices P y W se definieron en las ecuaciones (17) y (18), respectivamente,

$$P = \begin{pmatrix} X & 0 \\ 0 & \mathbf{1} \end{pmatrix} \quad \text{y} \quad W = \begin{pmatrix} W_{\beta\beta} & W_{\beta\phi} \\ W_{\phi\beta} & W_{\phi\phi} \end{pmatrix}$$

con

$W_{\beta\beta} = diag\left(\phi^2 a_i \left(\frac{d\mu_i}{d\eta_i}\right)\right)$, $\quad W_{\beta\phi} = diag\left(\phi\{\mu_i a_i - \psi'((1-\mu_i)\phi)\}\left(\frac{d\mu_i}{d\eta_i}\right)\right)$ y $W_{\phi\phi} = diag(b_i)$

donde a_i y b_i fueron definidas en la sección 1.3. Por lo tanto, la matriz de información de Fisher para el vector de parámetros $\theta = (\beta^t, \phi)^t$ es $K(\theta) = P^t W P$.

Por medio de cálculos simples es posible concluir que las ecuaciones (20) y la matriz de información de Fisher están de acuerdo con las expresiones del vector score y la la matriz de información de Fisher presentadas en Ferrari & Cribari-Neto (2004).

Capítulo 2
Inferencia y validación de supuestos

A continuación se realiza el proceso de inferencia del modelo propuesto, al igual que la validación de algunos supuestos deseables en un modelo lineal de regresión tradicional, resaltando que no estricto el cumplimiento de estos supuestos en el modelo de regresión beta con distancias, ya que el método de distancias lo hace no lineal en la variables explicativas y por el lado, la variable beta sería no lineal en la respuesta.

2.1. Inferencia para muestras grandes

A continuación se hace inferencia en muestras grandes desarrollando las pruebas de razón de verosimilitud, de score y Wald para los parámetros de regresión beta con distancias. También se obtienen los intervalos de confianza para la precisión y los parámetros de regresión, asumiendo como función de enlace la logit.

En el modelo de regresión beta con distancias tiene interés la siguiente prueba de hipótesis

$$H_0 : \theta_* = (\beta_*^t, \alpha_*^t)^t = \theta_*^{(0)} \quad versus \quad H_1 : \theta_* = (\beta_*^t, \alpha_*^t)^t \neq \theta_*^{(0)} \tag{21}$$

donde $\theta_* = (\beta_1, \ldots, \beta_m, \alpha_1, \ldots, \alpha_b)^t$ y $\beta_*^{(0)} = (\beta_1^{(0)}, \ldots, \beta_m^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_b^{(0)})^t$ para $m < k$ y $b < h$ dados.

La estadística de razón de log-verosimilitud para juzgar (21) está dada por

$$w_1 = 2\left(l\left(\widehat{\beta}, \widehat{\alpha}\right) - l\left(\widetilde{\beta}, \widetilde{\alpha}\right)\right) \tag{22}$$

donde $l(\widehat{\beta}, \widehat{\alpha})$ es la función log-verosímil y $(\widetilde{\beta}^t, \widetilde{\alpha}^t)^t$ es el estimador máximo verosímil restringido de $(\beta^t, \alpha^t)^t$ bajo la hipótesis nula. Bajo las condiciones usuales de regularidad y bajo H_0, $w_1 \xrightarrow{d} \chi^2_{(m+b)}$; de este modo, el juzgamiento de (21) puede llevarse a cabo usando los valores críticos aproximados de la distribución asintótica $\chi^2_{(m+b)}$.

En especial para contrastar $H_0 : \beta_* = \beta_*^{(0)}$ contra $H_1 : \beta_* \neq \beta_*^{(0)}$, se describe el estadístico Score, sea $U_{1\beta}$ el vector de dimensión m que contiene los primeros m elementos de la función Score para β y sea $K_{11}^{\beta\beta}$ la matriz de dimensiones $(m \times m)$

formada por las primeras m filas y las primeras m columnas de $K^{\beta\beta}$. De la ecuación (16) se puede mostrar que $U_{1\beta} = X_1^t \Upsilon T_1 (y^* - \mu^*)$, donde X es particionada como $[X_1 \vdots X_2]$ siguiendo la partición de β.

De lo anterior, la estadística Score de Rao para contrastar $H_0 : \beta_* = \beta_*^{(0)}$ está dada por

$$w_2 = \widetilde{U}_{1\beta}^t \widetilde{K}_{11}^{\beta\beta} \widetilde{U}_{1\beta} \tag{23}$$

donde las tildes indican que las cantidades son evaluadas en el estimador máximo verosímil restringido. Bajo las condiciones usuales de regularidad y bajo H_0, $w_2 \xrightarrow{d} \chi^2_{(m)}$.

La inferencia asintótica también puede ser llevada a cabo usando la estadística de Wald, la cual está dada por

$$w_3 = \left(\widehat{\beta}_* - \beta_*^{(0)}\right)^t \left(\widehat{K}_{11}^{\beta\beta}\right)^{-1} \left(\widehat{\beta}_* - \beta_*^{(0)}\right) \tag{24}$$

donde $\widehat{K}_{11}^{\beta\beta}$ es igual a $K_{11}^{\beta\beta}$ evaluada en el estimador máximo verosímil sin restricción, y $\widehat{\beta}_*$ es el estimador máximo verosímil de β_*. Bajo las condiciones usuales de regularidad y bajo H_0, $w_3 \xrightarrow{d} \chi^2_{(m)}$.

Particularmente, para juzgar la significancia del j-ésimo parámetro de regresión β_j, $j = 1, \ldots, k$ se puede usar la raíz cuadrada positiva de la estadística de Wald, es decir, si se desea contrastar la hipótesis

$$H_0 : \beta_j = 0 \quad \textit{versus} \quad H_1 : \beta_j \neq 0$$

La estadística de prueba será entonces

$$z = \frac{\widehat{\beta}_j}{ee(\widehat{\beta}_j)} \tag{25}$$

donde $ee(\widehat{\beta}_j)$ es el error estándar asintótico del estimador máximo verosímil de $\widehat{\beta}_j$ obtenido de la inversa de la matriz de información de Fisher evaluada en las estimaciones máximo verosímiles. El límite de la distribución de la estadística (25) bajo H_0 cierta es una normal estándar.

Adicionalmente, un intervalo aproximado del $(1-q)100\,\%$ de confianza para β_j, $j = 1, \ldots, k$ y $0 < q < 1/2$ está dado por

$$IC_{(1-q)100\,\%}(\beta_j) = \left(\widehat{\beta}_j - \Phi^{-1}_{(1-q/2)} ee\left(\widehat{\beta}_j\right); \widehat{\beta}_j + \Phi^{-1}_{(1-q/2)} ee\left(\widehat{\beta}_j\right)\right)$$

donde Φ^{-1} es la inversa de la distribución de una normal estándar.

Si se desean calcular regiones de confianza aproximadas para grupos de parámetros de regresión, éstas pueden ser obtenidas al invertir alguna de las tres pruebas para muestras grandes presentadas en (22), (23) y (24).

De forma similar, para juzgar $H_0 : \alpha_* = \alpha_*^{(0)}$ contra $H_1 : \alpha_* \neq \alpha_*^{(0)}$ se decribe el estadístico Score, sea $U_{1\alpha}$ el vector de dimensión b que contiene los primeros b elementos de la función Score para α. Además, sea $K_{11}^{\alpha\alpha}$ la matriz de dimensión $(b \times b)$ formada por las primeras b filas y las primeras b columnas de $K^{\alpha\alpha}$. De

este modo, de la ecuación (16) se puede mostrar que $U_{1\alpha} = Z_2^t T_2 v$, donde Z es particionada como $[Z_1 \vdots Z_2]$ siguiendo la partición de α.

Por lo tanto, la estadística Score para juzgar $H_0 : \alpha_* = \alpha_*^{(0)}$ está dada por

$$w_4 = \widetilde{U}_{1\alpha}^t \widetilde{K}_{11}^{\alpha\alpha} \widetilde{U}_{1\alpha}$$

donde las tildes indican que las cantidades son evaluadas en el estimador máximo verosímil restringido. Bajo las condiciones usuales de regularidad y bajo H_0, $w_4 \xrightarrow{d} \chi^2_{(b)}$.

La inferencia asintótica también puede ser llevada a cabo usando la estadística de Wald, la cual está dada por

$$w_5 = \left(\widehat{\alpha}_* - \alpha_*^{(0)}\right)^t \left(\widehat{K}_{11}^{\alpha\alpha}\right)^{-1} \left(\widehat{\alpha}_* - \alpha_*^{(0)}\right)$$

donde $\widehat{K}_{11}^{\alpha\alpha}$ es igual a $K_{11}^{\alpha\alpha}$ evaluada en el estimador máximo verosímil sin restricción, y $\widehat{\alpha}_*$ es el estimador máximo verosímil de α_*. Bajo las condiciones usuales de regularidad y bajo H_0, $w_5 \xrightarrow{d} \chi^2_{(b)}$.

En este caso, para juzgar la significancia del s-ésimo parámetro de regresión α_s, $s = 1, \ldots, h$ se puede usar la raíz cuadrada positiva de la estadística de Wald, es decir, si se desea contrastar la hipótesis

$$H_0 : \alpha_s = 0 \quad \textit{versus} \quad H_1 : \alpha_s \neq 0 \tag{26}$$

La estadística de prueba será entonces

$$z = \frac{\widehat{\alpha}_s}{ee(\widehat{\alpha}_s)} \tag{27}$$

donde $ee(\widehat{\alpha}_s)$ es el error estándar asintótico del estimador máximo verosímil de $\widehat{\beta}_j$ obtenido de la inversa de la matriz de información de Fisher evaluada en los estimaciones máximo verosímiles. El límite de la distribución de la estadística (27) bajo H_0 cierta es una normal estándar.

Adicionalmente, un intervalo aproximado del $(1-q)100\,\%$ de confianza para α_s, $s = 1, \ldots, h$ y $0 < q < 1/2$ está dado por

$$IC_{(1-q)100\,\%}(\alpha_s) = \left(\widehat{\alpha}_s - \Phi^{-1}_{(1-q/2)} ee\left(\widehat{\alpha}_s\right); \widehat{\alpha}_s + \Phi^{-1}_{(1-q/2)} ee\left(\widehat{\alpha}_s\right)\right) \tag{28}$$

donde Φ^{-1} es la inversa de la distribución de una normal estándar.

Al igual que antes, si se desean calcular regiones de confianza aproximadas para grupos de parámetros de regresión, éstas pueden ser obtenidas al invertir alguna de las tres pruebas para muestras grandes presentadas en (26), (27) y (28).

2.2. Predicción de un nuevo individuo

Si se supone que sobre las variables mixtas explicativas se ha observado un nuevo individuo $n+1$ del que se conoce las observaciones sobre las variables independientes $(v_{n+1} = (v_{(n+1)0}, v_{(n+1)1}, ..., v_{(n+1)p}))$; tales observaciones permiten calcular las

distancias entre el individuo $n+1$ y cada uno de los individuos que intervinieron en el modelo planteado en (5), es decir,

$$\delta_{(n+1)i} = \delta(v_{n+1}, v_i), \quad v_i \in \Omega, \quad i = 1, \ldots, n$$

A partir de estas distancias se puede hacer una predicción empleando el siguiente resultado Gower (1968), que relaciona el vector $d = (\delta^2_{(n+1)1}, \ldots, \delta^2_{(n+1)n})^t$ de los cuadrados de estas distancias con el vector $x_{n+1} = (x_{(n+1)1}, \ldots, x_{(n+1)p})$ de las coordenadas principales atribuibles al nuevo individuo.

$$\begin{aligned} \delta^2_{(n+1)i} =& (x_{n+1} - x_i)(x_{n+1} - x_i)^t \\ =& x_{n+1}x^t_{n+1} + x_i x^t_i - 2x_{n+1}x^t_i \end{aligned} \tag{29}$$

Sumando para i de 1 a n, y teniendo en cuenta que las columnas de la matriz X de coordenadas suman 0, se obtiene:

$$\sum_{i+1}^{n} \delta^2_{(n+1)i} = n x_{n+1} x^t_{n+1} + tr(B)$$

Sustituyendo esta última ecuación en (29), se tiene

$$2x_{n+1}x^t_i = \frac{1}{n}\left(\sum_{i=1}^{n} \delta^2_{(n+1)i} - tr(B)\right) + b_{ii} - \delta^2_{(n+1)i}$$

Al considerar las diferencias con los n individuos de forma matricial

$$2Xx^t_{n+1} = \frac{1}{n}\left(\sum_{i=1}^{n} \delta^2_{(n+1)i} - tr(B)\right) 1_n + (b - d)$$

Premultiplicando por X^t y dado que $X^t 1_n = 0$, se encuentra que

$$\begin{aligned} 2X^t X x^t_{n+1} =& X^t(b - d) \\ x^t_{n+1} =& \frac{1}{2}(X^t X)^{-1} X^t (b - d) \\ =& \frac{1}{2}\Lambda^{-1} X^t (b - d) \end{aligned}$$

La predicción es entonces

$$\widehat{\eta}_{(n+1)\beta} = \widehat{\beta}_0 + x^t_{n+1}\widehat{\beta}$$

Si se considera ahora el modelo DB en dimensión k y se hace la partición:

$$x_{n+1} = \begin{pmatrix} x_{(k)} \\ l \end{pmatrix}, \quad X = \begin{pmatrix} X_{(k)} & L \end{pmatrix} \quad \text{y} \quad \Lambda = \begin{pmatrix} \Lambda_k & 0 \\ 0 & \Lambda_{m-k} \end{pmatrix}$$

donde $x_{(k)} = (x_1, \ldots, x_k)^t$ son las k coordenadas relativas de las k-dimensiones predictivas asociadas al $n+1$ individuo, y la diagonal Λ_k contiene los valores propios, así se obtiene:

$$\widehat{\eta}_{(n+1)\beta} = \widehat{\beta}_0 + x_{(k)}^t \widehat{\beta} + l^t \widehat{\beta}$$

como l contiene las coordenadas menos correlaccionadas en el nuevo individuo $n+1$, entonces

$$\widehat{\eta}_{(n+1)\beta}(k) = \widehat{\beta}_0 + x_{(k)}^t \widehat{\beta} \tag{30}$$

Obsérvese que si l es muy grande, entonces (30) no funciona bien y entonces la observación v_{n+1} podría ser un atípico.

Finalmente, un intervalo aproximado del $(1-q)100\,\%$ de confianza para la media de la respuesta μ, para un conjunto de valores v_{n+1} dado, puede ser calculado a través de

$$\left(g^{-1}\left(\widehat{\eta}_{(n+1)\beta}(k) - \Phi^{-1}_{(1-q/2)} ee(\widehat{\eta}_{(n+1)\beta}(k) \right);\ g^{-1}\left(\widehat{\eta}_{(n+1)\beta}(k) + \Phi^{-1}_{(1-q/2)} ee(\widehat{\eta}_{(n+1)\beta}(k) \right) \right)$$

donde $ee(\widehat{\eta}_{(n+1)\beta}(k)) = \sqrt{x_{(k)}^t \widehat{cov}(\widehat{\beta}) x_{(k)}}$, con $\widehat{cov}(\widehat{\beta})$ obtenida de la inversa de la matriz de información de Fisher evaluada en las estimaciones máximo verosímiles. Es de notar que en general, el intervalo anterior es válido para funciones de enlace estrictamente crecientes. El modelo (8) depende de la distancia δ_{ij} elegida, contiene el modelo de regresión beta como caso particular, pero es especialmente interesante en los casos de variables mixtas y no lineal, con tal de elegir una distancia adecuada.

Para el paramétro de presición asociado al $n+1$ individuo, realizando un procedimiento parecido al anterior, se encuentra que:

$$\widehat{\eta}_{(n+1)\alpha} = \widehat{\alpha}_0 + z_{(k)}^t \widehat{\alpha} + l^t \widehat{\alpha}$$

como l contiene las coordenadas menos correlaccionadas en el nuevo individuo $n+1$, entonces

$$\widehat{\eta}_{(n+1)\alpha}(k) = \widehat{\alpha}_0 + z_{(k)}^t \widehat{\alpha} \tag{31}$$

Al igual que para la media, obsérvese que si l es muy grande, entonces (31) no funciona bien y entonces la observación u_{n+1} podría ser un atípico.

Por último, un intervalo aproximado del $(1-q)100\,\%$ de confianza para la media de la respuesta ϕ, para un conjunto de valores u_{n+1} dado, puede ser calculado a través de

$$\left(g^{-1}\left(\widehat{\eta}_{(n+1)\alpha}(k) - \Phi^{-1}_{(1-q/2)} ee(\widehat{\eta}_{(n+1)\alpha}(k)\right);\ g^{-1}\left(\widehat{\eta}_{(n+1)\alpha}(k) + \Phi^{-1}_{(1-q/2)} ee(\widehat{\eta}_{(n+1)\alpha}(k)\right)\right)$$

donde $ee(\widehat{\eta}_{(n+1)\alpha}(k)) = \sqrt{z^t_{(k)}\widehat{cov}(\widehat{\alpha})z_{(k)}}$, con $\widehat{cov}(\widehat{\alpha})$ obtenida de la inversa de la matriz de información de Fisher evaluada en las estimaciones máximo verosímiles.

2.3. Medidas de diagnóstico

Después de ajustar el modelo de regresión beta de interés, es importante llevar a cabo un análisis diagnóstico, con el fin de verificar la bondad del ajuste del modelo estimado. Una medida global de la variación explicada se obtiene al calcular un pseudo R^2 definido como

$$R^2_p = r^2(\widehat{\gamma}, g(y)) \qquad 0 \leq R^2_p \leq 1$$

donde $r(\widehat{\gamma}, g(y))$ es el coeficiente de correlación muestral entre $\widehat{\gamma}$ y $g(y)$. Cuando $R^2_p = 1$ indica un acuerdo perfecto entre $\widehat{\gamma}$ y $g(y)$ y por lo tanto entre $\widehat{\mu}$ y y.

La discrepancia del modelo ajustado se puede determinar a través de qué tanto el modelo ajustado es significativamente diferente del modelo saturado (que contiene tantos parámetros como observaciones hay en el modelo n). Para ello, sea

$$D(y, \mu, \phi) = \sum_{i=1}^{n} 2(l_i(\widetilde{\mu}_i, \phi_i) - l_i(\mu_i, \phi_i))$$

donde $\widetilde{\mu}_i$ es el valor de μ_i que resulta de resolver $\partial l_i/\partial \mu_i = 0$, es decir, $\phi_i(y^*_{i^*} - \mu^*_i) = 0$. Cuando ϕ_i es grande, $\mu^*_i \approx \log\{\mu_i/(1-\mu_i)\}$ y de esto se tiene que $\widetilde{\mu}_i \approx y_i$.

Para un ϕ conocido esta medida de discrepancia entre los dos modelos es igual a $D(y; \bar{\mu}, \phi)$, donde $\bar{\mu}$ es el estimador máximo verosímil de μ bajo el modelo que está siendo estudiado. Cuando ϕ es desconocido, una aproximación a esta cantidad es

$$D(y; \widehat{\mu}, \widehat{\phi}) = \sum_{i=1}^{n} (r^d_i)^2 \tag{32}$$

la cual es conocida como la *devianza* y con

$$r^d_i = signo(y_i - \widehat{\mu}_i)\{2(l_i(\widetilde{\mu}_i, \widehat{\phi}_i) - l_i(\widehat{\mu}_i, \widehat{\phi}_i))\}^{1/2}$$

donde $l_i(\widetilde{\mu}_i, \widehat{\phi}_i)$ es la máxima verosimilitud del modelo saturado y $l_i(\widehat{\mu}_i, \widehat{\phi}_i)$ la máxima verosimilitud del modelo restringido bajo H_0. Como es de esperarse la log verosimilitud asociada al modelo saturado puede ser mayor que la asociada a un modelo con $m + h < n$ parámetros.

La estadística presentada en (32) tiene asintóticamente una distribución $\chi^2_{(n-m)}$. Claramente es deseable obtener una devianza pequeña, y si esta resulta no significativa, la conclusión es que el desempeño del modelo en estudio no es significativamente diferente del modelo saturado y por lo tanto, la estimación de los otros parámetros involucrados en el modelo saturado resulta innecesaria.

Por otra parte, debido a que la i-ésima observación contribuye una cantidad $(r_i^d)^2$ a la deviance, una observación con un alto valor absoluto de r_i^d puede ser atípica. De lo anterior, a r_i^d se le llama el i-ésimo residual de deviance.

Hay varias clases de residuos para modelos de regresión beta, una alternativa natural son los residuales de Pearson, los cuales Ferrari & Cribari-Neto (2004) llama residuales ordinarios estandarizados y se definen como

$$r_{P,i} = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{Var}(y_i)}}$$

donde $\widehat{Var}(y_i) = \widehat{\mu}_i(1-\widehat{\mu}_i)/(1+\widehat{\phi}_i)$, $\widehat{\mu}_i = g_1^{-1}(x_i^t\widehat{\beta})$ y $\widehat{\phi}_i = g_2^{-1}(z_i^t\widehat{\alpha})$. Similarmente, se pueden definir los residuales de deviance como se hizo en la ecuación (32) vía las contribuciones del signo al exceso de la verosimilitud. Además, Espinheira, Ferrari & Cribari-Neto (2008) propusieron unos residuales con mejores propiedades que los residuales de Pearson, éstos son llamados los residuales estandarizados ponderados 2:

$$r_{SW2,i} = \frac{y_i^* - \widehat{\mu}_i^*}{\sqrt{\widehat{m}_i(1-h_{ii})}}$$

donde $y_i^* = log\{y_i/(1-y_i)\}$ y $\mu_i^* = \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i)$, $\psi(\cdot)$ denota la función digamma. Además $m_i = \{\psi'(\mu_i\phi_i) + \psi'((1-\mu_i)\phi_i)\}$ y h_{ii} es el i-ésimo elemento de la matriz hat (ver Ferrari & Cribari-Neto (2004) y Espinheira et al. (2008)).

Un gráfico de estos residuales contra las observaciones indexadas (i) puede evidenciar patrones no detectables. Adicionalmente, una tendencia detectable en el gráfico de r_i o $r_{P,i}$ o $r_{SW2,i}$ contra $\widehat{\eta}_{1i}$ puede sugerir una falta de especificación de la función de enlace.

Dado que la distribución de los residuales no se conoce, los gráficos Half-Normal con valores de cubiertas simuladas son una buena estrategia de diagnóstico, véase Atkinson (1985) y Neter, Kutner, Nachtsheim & Wasserman (1996) para mayores detalles. La idea principal es aumentar el gráfico Half-Normal usual al adicionar una cubierta simulada, que es útil para decidir cuándo los residuales observados son consistentes con el modelo ajustado. Este tipo de gráficos se obtienen siguiendo los siguientes pasos:

(1) Ajuste el modelo y genere una muestra simulada de n observaciones independientes utilizando el modelo ajustado como si éste fuera el modelo verdadero.
(2) Ajuste un modelo con los datos de la muestra generada y calcule los valores absolutos ordenados de los residuales.
(3) Repita los pasos (1) y (2) k veces.
(4) Considere los n conjuntos de k estadísticas de orden, para cada conjunto calcule el promedio y los valores mínimo y máximo.

(5) Grafique los valores obtenidos y los residuales ordenados de la muestra original contra los puntajes Half-Normal de la forma: $\Phi^{-1}((t+n-1/8)/(2n+1/2))$.

Los valores mínimo y máximo de las k estadísticas de orden generan la cubierta. Atkinson (1985) sugiere usar $k = 19$, ya que la probabilidad de que un residual absoluto caiga más allá de la banda superior proporcionada por la cubierta es aproximadamente igual a $1/20 = 0{,}05$. Las observaciones correspondientes a los residuales absolutos que se encuentren fuera de los límites proporcionados por la cubierta simulada se deben estudiar a mayor profundidad ya que pueden ser observaciones atípicas o influyentes. Adicionalmente, si una proporción considerable de puntos cae fuera de la cubierta, se tiene evidencia en contra del adecuado ajuste del modelo propuesto.

Después de hacer una identificación de las observaciones influyentes y un análisis de residuales, se hace uso de la generalización de los leverage propuestos por Wei, Hu & Fung (1998), los cuales se definen como:

$$GL(\widetilde{\theta}) = \frac{\partial \widetilde{y}}{\partial y^t}$$

donde θ es un vector de dimensión s tal que $E(y) = \mu(\theta)$ y $\widetilde{\theta}$ es un estimador de θ, con $\widetilde{y} = \mu(\widetilde{\theta})$. De aquí, el elemento (i,u) de $GL(\widetilde{\theta})$, es decir el leverage generalizado del estimador $\widetilde{\theta}$ en (i,u), es la razón de cambio instantánea en el i-ésimo valor predicho con respecto al u-ésimo valor de la respuesta. Como se muestra en Ferrari y Cribari-Neto (2004), el leverage generalizado es invariante a la reparametrización, y las observaciones con un gran GL_{iu} son puntos leverage.

Sea $\widehat{\theta}$ el estimador máximo verosímil de θ; asumiendo que existe, es único y que la función log verosímil tiene derivadas continuas de segundo orden con respecto a θ y a y, Wei et al. (1998) mostraron que el leverage generalizado se obtiene al evaluar

$$GL(\theta) = D_\theta \left(-\frac{\partial^2 l}{\partial\theta\partial\theta^t} \right)^{-1} \frac{\partial^2 l}{\partial\theta\partial y^t}$$

en $\widehat{\theta}$, donde $D_\theta = \partial\mu/\partial\theta^t$.

Como primer paso, se puede obtener una forma cerrada para $GL(\beta)$ en el modelo de regresión beta de distancias propuesto, bajo la suposición que los ϕ's son conocidos. Se puede comprobar que $D_\beta = TX$ y se tiene que:

$$-\frac{\partial^2 l}{\partial\beta\partial\beta^t} = X^t Q X$$

donde $Q = diag\{q_1, \ldots, q_n\}$, con

$$q_i = \left(\phi_i^2\{\psi'(\mu_i\phi_i) + \psi'((1-\mu_i)\phi_i)\} + \phi_i(y_i^* - \mu_i^*)\frac{g''(\mu_i)}{g'(\mu_i)} \right) \frac{1}{\{g'(\mu_i)\}^2}, \qquad i = 1, \ldots, n$$

Adicionalmente, se puede mostrar que $\partial^2 l/\partial\beta\partial y^t = X^t T M$, siendo $M = diag\{m_1, \ldots, m_n\}$ con $m_i = \phi_i/\{y_i(1-y_i)\}$ para $i = 1, \ldots, n$. De lo anterior, se tiene que

$$GL(\beta) = TX(X^t Q X)^{-1} X^t T M \tag{33}$$

Al reemplazar la información observada, $-\partial^2 l/\partial\beta\partial\beta^t$, por la información esperada, $E(-\partial^2 l/\partial\beta\partial\beta^t)$, la expresión para $GL(\beta)$ es dada por (33) reemplazando Q por W obteniendo $GL^*(\beta)$. Es de destacar que los elementos diagonales de $GL^*(\beta)$ son los mismos que los de $M^{1/2}TX(X^tWX)^{-1}X^tM^{1/2}$, y que $M^{1/2}T$ es una matriz diagonal cuyo i-ésimo elemento de la diagonal es $\{g'(\mu_i)V(y_i)^{1/2}\}^{-1}$. También es importante señalar que existe una estrecha relación entre los elementos diagonales de $GL^*(\beta)$ y los de la matriz "sombrero" de costumbre,

$$H = W^{1/2}X(X^tWX)^{-1}X^tW^{1/2}$$

cuando el mín$\{\phi_1, ..., \phi_n\}$ es grande. La relación se basa en que cuando mín$\{\phi_1, ..., \phi_n\}$ es grande, el i-ésimo elemento de la diagonal de $W^{1/2}$ es aproximadamente igual a $\{g'(\mu_i)V(y_i)^{1/2}\}^{-1}$.

Ahora considérese la situación cuando los ϕ_i's son desconocidos, de aquí $\theta^t = (\beta^t, \alpha^t)$. Así, $D_\theta = [TX \quad 0]$ donde 0 es una matriz de $n \times h$ ceros. También $-\partial^2 l/\partial\theta\partial\theta^t$ se presento en (19) con W reemplazado por Q. Es claro que la inversa de $-\partial^2 l/\partial\theta\partial\theta^t$ es dada por la expresión (19) con W reemplazada por Q. Adicionalmente,

$$\frac{\partial^2 l}{\partial\theta\partial y^t} = \begin{pmatrix} X^tTM \\ u^t \end{pmatrix}$$

donde $u = (u_1, \ldots, u_n)^t$ con $u_i = -(y_i - \mu_i)/\{y_i(1 - y_i)\}$, $i = 1, \ldots, n$. Se puede entonces demostrar que

$$GL(\beta, \alpha) = GL(\beta) + f(\alpha, \beta)$$

donde $GL(\beta)$ fue presentado en (33) y $f(\alpha, \beta)$ es una función matricial que depende de α y β. Cuando los diferentes ϕ_i's son grandes, $GL(\beta, \alpha) \approx GL(\beta)$.

Una medida de influencia de cada observación sobre los parámetros de regresión estimados es la distancia de Cook (Cook 1977) dada por $k^{-1}\left(\widehat{\beta} - \widehat{\beta}_{(i)}\right)^t X^tWX\left(\widehat{\beta} - \widehat{\beta}_{(i)}\right)$, donde $\widehat{\beta}_{(i)}$ es el parámetro estimado sin la i-ésima observación. Esta médida es la distancia al cuadrado entre $\widehat{\beta}$ y $\widehat{\beta}_{(i)}$. Para evitar el ajuste del modelo $n + 1$ veces, se utilizará la aproximación usual de la distancia de Cook dada por

$$C_i = \frac{h_{ii}r_i^2}{k(1 - h_{ii})^2}$$

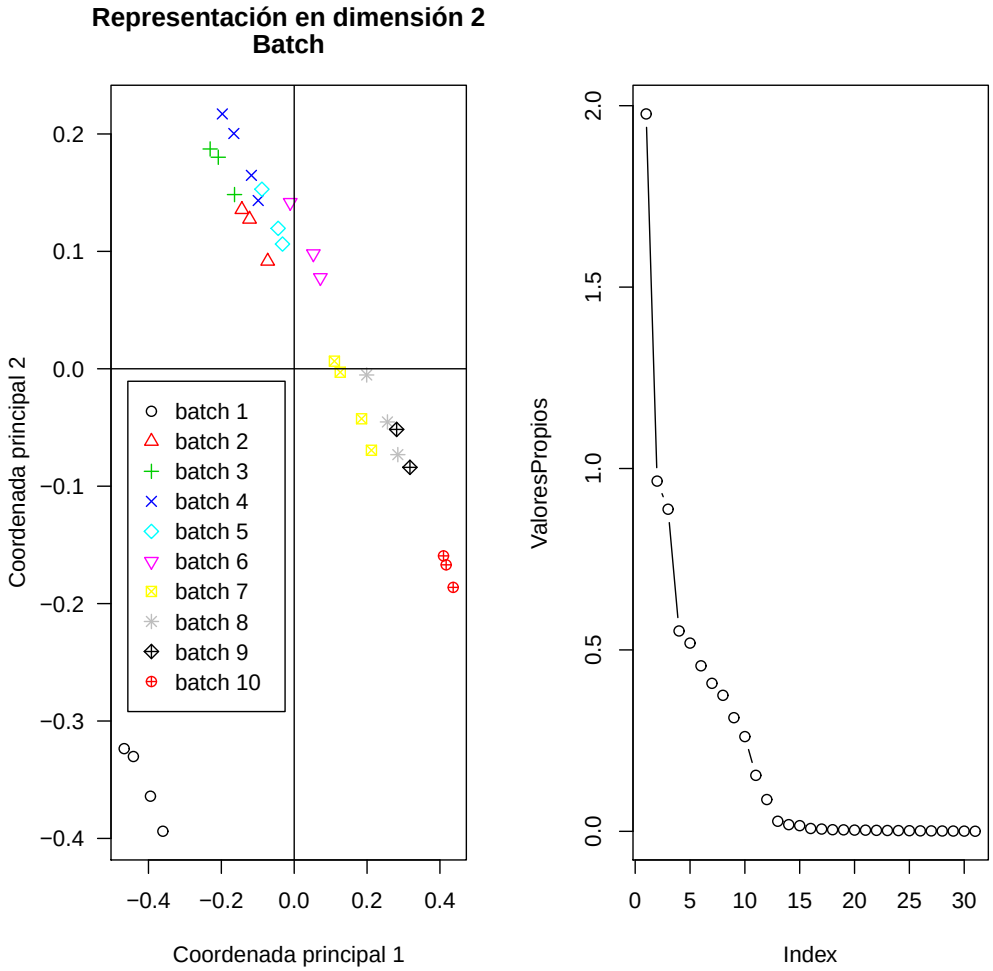
Esta expresión combina leverage y residuales. Además es común hacer un gráfico de C_i contra i para verificar posibles observaciones influyentes.

Finalmente, se pueden utilizar otras medidas de diagnóstico, tales como las medidas de influencia local (Cook 1986).

Capítulo 3
Aplicación

En esta aplicación se considera un modelo de regresión beta de distancias con precisión constante y variable. A continuación se consideran los datos de petróleo convertido a gasolina recolectados por Prater (1956). Los datos contienen treinta y dos observaciones con cinco variables. Se desea modelar la proporción de crudo convertido en gasolina tras un proceso de destilación y las covariables son: la gravedad del crudo (grados API), la presión del vapor del petróleo (lbf/in^2), la temperatura en la cual el 10 % se convierte en vapor y la temperatura (^oF). Existen diez conjuntos de crudos diferentes sujetos a condiciones controladas de destilación. Los datos están ordenados en forma ascendente de acuerdo con la covariable que mide la temperatura a la cual el 10 % del petróleo pasa a ser vapor. Esta variable asume diez diferentes valores y se utilizan para definir diez lotes de petróleo. La relación final está determinada por nueve variables dummy para los primeros nueve lotes de petróleo y la covariable temperatura (^oF) en la cual la gasolina se evapora. Este conjunto de datos fue analizada por Atkinson (1985), quien uso el modelo de regresión lineal y encontró que la distribución de los errores no era totalmente simétrica. Después, Ferrari & Cribari-Neto (2004), Ospina et al. (2006) y Simas et al. (2010) utilizaron estos datos como una ilustración del modelo de regresión beta, los esquemas de corrección del sesgo y de dispersión variable, respectivamente.

A partir del enfoque planteado en esta aplicación, se ajustan los modelos de regresión beta de distancias con precisión constante y variable a partir del uso de la distancia de Gower ya que las variables explicativas son mixtas. En primer lugar se construyen las distancias a partir de las variables explicativas y a continuación se construye la matriz $B = \left(I - \frac{1}{32}11^t\right) A \left(I - \frac{1}{32}11^t\right)$, a partir de la cual se seleccionan las primeras 12 componentes al hacer descomposición espectral, recogiéndose el 98.5 % de la variabilidad total en estas nuevas variables, como se puede observar en la gráfica 1 (gráfico de la derecha). Además, en la gráfica de la izquierda se observan las diez clases de crudo en la proyección de la información en dos dimensiones, en la parte derecha de este gráfico se encuentra crudos con mayor densidad, mientras en la izquierda con menos densidad. Los crudos de la parte superior tienen se pueden destilar con mayor facilidad y los de la parte inferior no lo son. Es importante notar que los gráficos presentados aquí se pueden obtener al ejecutar el programa en R presentado en el Apéndice, en éste también se obtienen los modelos presentados en las siguientes secciones.



Gráfica 1. Proyección en el espacio bidimensional y valores propios.

3.1. Modelo de regresión beta con distancias y precisión constante

Una vez realizado el anterior procedimiento se ajusta un modelo con precisión constante, utilizando diferentes funciones de enlace; los resultados para el pseudo-coeficiente de correlación, la estimación del parámetro de precisión constante y el valor del AIC para cada una de las funciones de enlace se encuentran en la tabla 1. Como se observa la función de enlace que aumenta el valor del pseudo R^2 es el log – log, pero su principal ventaja con respecto a las otras dos funciones de enlace es el aumento significativo de la estimación del parámetro de precisión, ya que es el más alto (758,1); además, el AIC es el más pequeño (-158.56). Por tal razón, se

presenta a continuación el modelo obtenido a partir del ajuste de regresión beta con distancias a través del uso de la función de enlace $\log - \log$.

TABLA 1. Funciones de enlace utilizadas en el modelo de regresión beta con distancias y precisión constante para el ajuste de la proporción de crudo convertido a gasolina.

Función de enlace	Pseudo - R^2	AIC	Párametro de precisión
logit	0.9596	-138.48	399.88
probit	0.9727	-147.07	525.70
cloglog	0.9474	-130.52	310.91
loglog	0.9823	-158.56	758.10

Al utilizar la función de enlace $\log - \log$, la hipótesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{12} = 0$ se rechaza ya que la prueba de razón de verosimilitud $\chi^2 = 91{,}93 > \chi^2_{(14,0,05)}$ (valor p$< 0{,}05$), con lo cual se concluye que existe al menos un término diferente de cero y que hay relación entre las variables explicativas con el porcentaje de crudo convertido a gasolina. En la tabla 2 se presentan los coeficientes de la curva ajustada con sus respectivas pruebas de significancia, en esta tabla se puede observar que todos las componentes son significativos en el modelo al nivel del 5 %, es de notar que se excluyeron las componentes X7 y X11 ya que no eran significativas. Además, el parámetro de precisión también es significativo al 5 %.

TABLA 2. Coeficientes estimados para la regresión beta con distancia y precisión constante que relaciona el porcentaje de crudo con las nuevas variables.

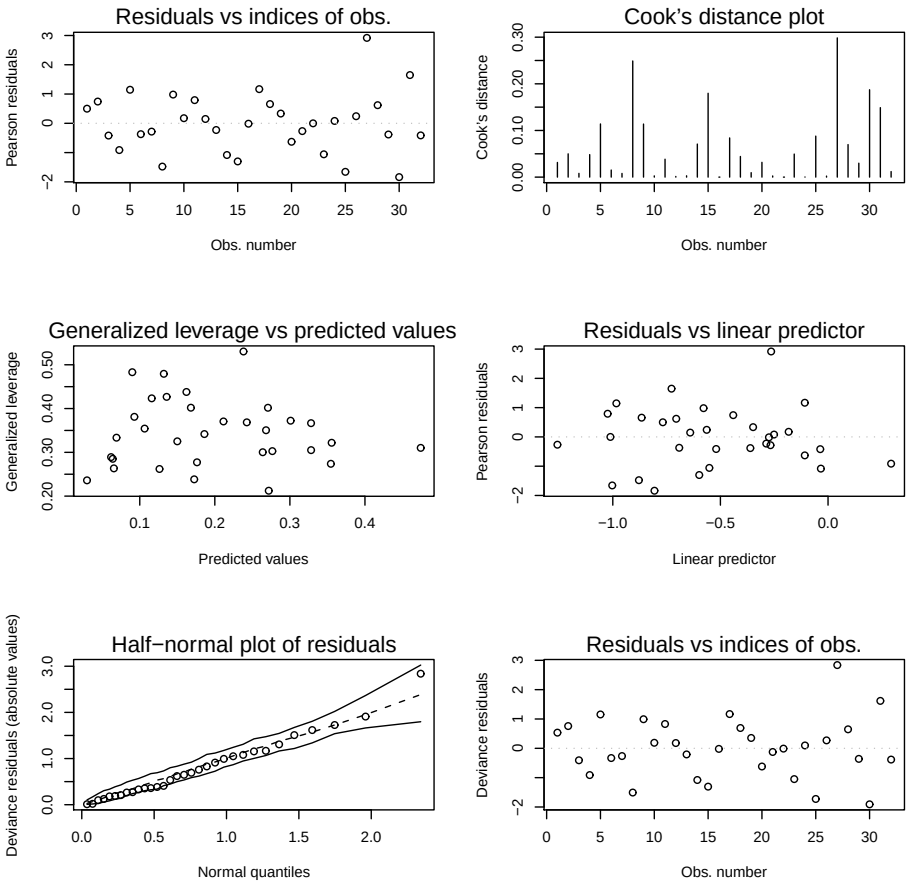
Efecto	Coeficiente	Error Est.	z	Valor $Pr(>\|z\|)$
Coeficientes (modelo medio con enlace log-log)				
(Intercept)	-0.5263	0.0086	-60.760	0.000
X1	-0.2819	0.0345	-8.166	0.000
X2	-0.4709	0.0500	-9.406	0.000
X3	-1.0201	0.0537	-18.990	0.000
X4	-1.5479	0.0655	-23.610	0.000
X5	-0.5764	0.0675	-8.535	0.000
X6	0.6745	0.0729	9.242	0.000
X8	0.1969	0.0792	2.485	0.013
X9	-0.9843	0.0862	-11.410	0.000
X10	1.5175	0.0943	16.090	0.000
X12	0.7963	0.1649	4.828	0.000
Coeficiente Phi (modelo de precisión con enlace identidad)				
(phi)	696.9	174.2	4.001	0.000

Así el modelo obtenido en términos de las componentes es:

$$-\log(-\log(\widehat{\mu})) = -0{,}53 - 0{,}28X1 - 0{,}47X2 - 1{,}02X3 - 1{,}54X4 - 0{,}57X5 + 0{,}67X6 + 0{,}19X8 - 0{,}98X9 + 1{,}52X10 + 0{,}79X12$$

Después del ajuste del modelo, es muy importante analizar la bondad de ajuste del modelo estimado. Como se menciono anteriormente el modelo cuenta con un

pseudo R^2 de 0.9823. Dado que la distribución de los residuales no es conocida, el half-normal plot simulado es un útil instrumento de diagnóstico, su interpretación es similar a un qq-plot, de manera que su uso permite decidir cuáles de los residuales observados son consistentes con el modelo ajustado. En la gráfica 2 se observa un ajuste adecuado del presente modelo, ya que no hay observaciones influyentes según el gráfico de los Cook's y tampoco hay observaciones atípicas que se salgan del intervalo -3 y 3 de acuerdo a los residuos de Pearson y de deviance, aunque la observación 27 tiene un valor un poco alto. Adicionalmente, en el gráfico de normalidad no se evidencia problemas ya que los residuales de deviance quedan dentro de los intervalos de confianza del 95 %.



Gráfica 2. Diagnóstico para el modelo con precisión constante.

3.2. Modelo de regresión beta con distancias y precisión variable

Ahora se ajusta un modelo con precisión variable utilizando diferentes funciones de enlace. Los resultados del pseudo-coeficiente de correlación y el valor del AIC para cada una de las funciones de enlace se encuentran en la tabla 3. Como se observa la función de enlace que aumenta el valor del pseudo R^2 (97.48 %) es el $\log - \log$ y el AIC es el más pequeño (-284.63). Por tal razón, al igual que en el modelo con precisión constante se ajusta el modelo de regresión beta con distancias a través del uso de la función de enlace $\log - \log$.

TABLA 3. Funciones de enlace utilizadas en el modelo de regresión beta con distancias y precisión variable para el ajuste de la proporción de crudo convertido a gasolina.

Función de enlace	Pseudo - R^2	AIC
logit	0.9388	-271.24
probit	0.9578	-280.01
cloglog	0.9242	-259.20
loglog	0.9748	-284.63

En este caso, en primer lugar se contrasta la hipótesis $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_{12} = \phi$ (precisión constante) utilizando como función de enlace log para la precisión, para lo cual se ajusta el modelo relacionando la gravedad del crudo, la presión del vapor del petróleo, la temperatura en la cual el 10 % se convierte en vapor, la temperatura (^oF) y los diez conjuntos de crudos con los parámetros de precisión. A partir de esto, se encuentra que al comparar con el modelo de regresión beta con distancias y precisión constante, que al nivel de significancia del 5 % se rechaza la hipótesis de precisión constante ($\chi^2 = 150{,}07 > \chi^2_{(12,0,05)}$, valor p= $2{,}2e-16 < 0{,}05$) y se concluye que la precisión es variable.

Al utilizar la función de enlace $\log - \log$, la hipótesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{12} = 0$ se rechaza al nivel de significancia del 5 % ya que la prueba de razón de verosimilitud $\chi^2 = 168{,}3 > \chi^2_{(26,0,05)}$ (valor p$< 0{,}05$), con lo cual se concluye que existe al menos un término diferente de cero y que hay relación entre las componentes con el porcentaje de crudo convertido a gasolina. En la tabla 4 se presentan los coeficientes con sus respectivas pruebas de significancia tanto para el modelo de la media como para el modelo de precisión variable.

En la tabla 4, se puede observar que todas las componentes en el modelo de media son significativos al 5 %, mientras que aparentemente en el modelo de presición variable, las componentes X7 y X9 no lo son; sin embargo, al realizar la prueba de razón de verosimilitud con la finalidad de revisar la exclusión de los dos términos se encuentra que su inclusión en el modelo de precisión es significativa al 5 % ($\chi^2 = 45{,}816 > \chi^2_{(2,0,05)}$, valor p= 0), por lo cual se incluyen.

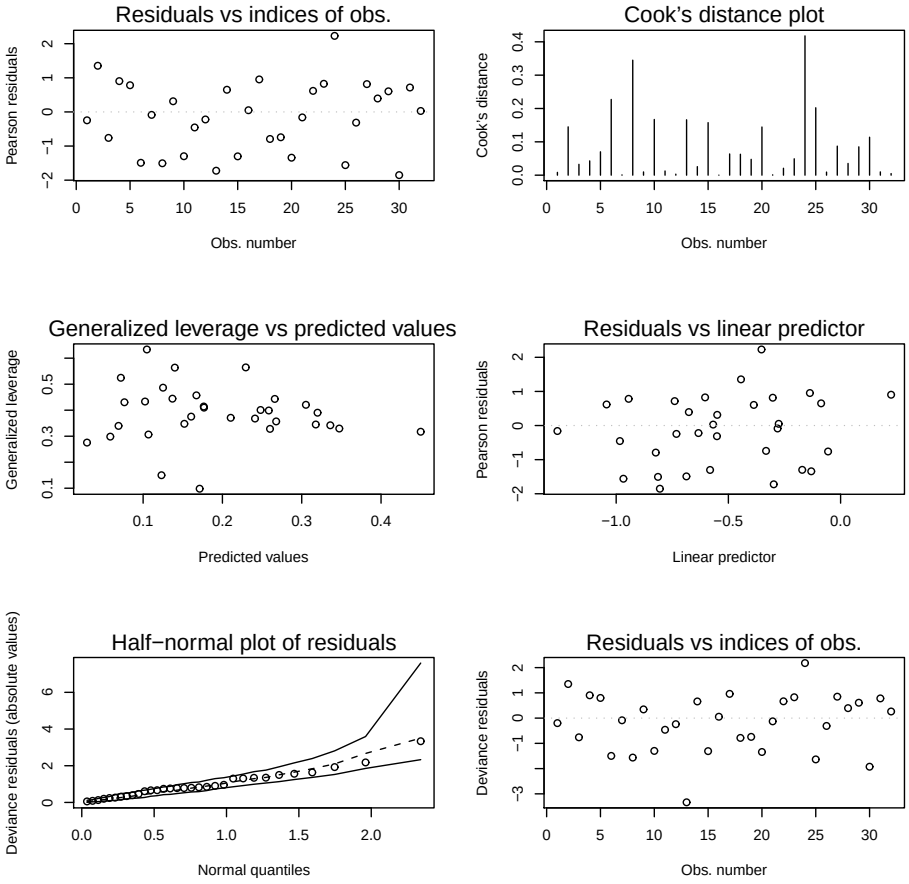
TABLA 4. Coeficientes estimados para la regresión beta con distancia y precisión variable que relaciona el porcentaje de crudo con las nuevas variables.

Efecto	Coeficiente	Error Est.	z	Valor $Pr(>\|z\|)$
Coeficientes (modelo medio con enlace $\log - \log$)				
(Intercept)	-0.5309	0.0046	-113.261	0.000
X1	-0.3210	0.0289	-11.096	0.000
X2	-0.3923	0.0250	-15.636	0.000
X3	-1.0273	0.0242	-42.430	0.000
X4	-1.3925	0.0409	-34.008	0.000
X5	-0.5422	0.0302	-17.898	0.000
X6	0.6544	0.0339	19.284	0.000
X7	0.1080	0.0457	2.360	0.018
X8	0.0939	0.0175	5.343	0.000
X9	-0.9570	0.0333	-28.695	0.000
X10	1.3690	0.0184	74.370	0.000
X11	-0.0811	0.0049	-16.403	0.000
X12	0.6145	0.0069	87.859	0.000
Coeficiente Phi (modelo de precisión con enlace identidad)				
(Intercept)	11.1407	0.2418	46.079	0.000
X1	-8.7075	0.9689	-8.987	0.000
X2	17.9111	1.3985	12.807	0.000
X3	-13.6098	1.4536	-9.363	0.000
X4	-18.1602	1.8292	-9.928	0.000
X5	23.5601	1.8593	12.672	0.000
X6	-11.7314	2.0727	-5.660	0.000
X7	-1.9030	2.1601	-0.881	0.378
X8	38.9208	2.2398	17.377	0.000
X9	-3.3100	2.5160	-1.316	0.188
X10	-5.6950	2.7571	-2.066	0.039
X11	-39.5832	3.4389	-11.510	0.000
X12	-20.3000	4.6624	-4.354	0.000

Así el modelo obtenido en términos de las componentes para la media es:

$$-\log(-\log(\widehat{\mu})) = -0{,}53 - 0{,}32X1 - 0{,}39X2 - 1{,}03X3 - 1{,}39X4 - 0{,}54X5 + 0{,}65X6 + 0{,}10X7 + 0{,}09X8 - 0{,}96X9 + 1{,}37X10 - 0{,}08X11 + 0{,}61X12$$

Una vez ajustado el modelo es muy importante analizar la bondad de ajuste del modelo estimado. Como se menciono anteriormente el modelo cuenta con un pseudo R^2 de 0.9748. En la gráfica 3 se observa un ajuste adecuado del presente modelo, ya que no hay observaciones influyentes según el gráfico de los Cook's y tampoco hay observaciones atípicas que se salgan del intervalo -3 y 3 de acuerdo a los residuos de Pearson y de deviance, aunque la observación 24 tiene un valor un poco alto. Adicionalmente, en el gráfico de normalidad no se evidencia problemas ya que los residuales de deviance quedan dentro de los intervalos de confianza del 95 %.



Gráfica 3. Diagnóstico para el modelo con precisión variable.

Conclusiones

En este trabajo se propuso una metodología basada en distancias para ajustar variables respuesta tipo beta con precisión constante y variable; se ajusto el modelo de regresión beta de distancias, se obtuvo la estimación de los diferentes parámetros involucrados y se realizo la validación del modelo propuesto. Por medio de una aplicación se ilustro la metodología presentada en este trabajo, en dicha aplicación se ajustan los modelos de regresión beta de distancias con precisión constante y dispersión variable con covariables a partir del uso de la distancia de Gower.

Aunque para el caso de la aplicación se utilizo la distancia de Gower, se pueden emplear otras distancias, el método no tiene restricción en cuanto a la escogencia de una distancia en particular.

Por otro lado, como en la metodología propuesta se hace una transformación de la variable beta con la finalidad de utilizar las ideas del modelo lineal generalizado, se debe calcular el sesgo de los estimadores obtenidos utilizando un método como el bootstrap o el método de permutación de Fisher para obtener un mejor ajuste en los parámetros de precisión tanto constante como variable. Esto no fue realizado en este trabajo, pero puede ayudar a mejorar los intervalos de confianza que se hagan ya que esto no afecta mucho a los parámetros del modelo de la media, pero si al modelo de precisión.

Apéndice. Programa en R

```
# Lectura de Información

basepetro<-data('GasolineYield', package ='betareg')
edit(GasolineYield)
summary(GasolineYield)

# Libreria Betareg
library(betareg)

# Modelamiento sin distancias
# Modelo logit

gy_logit <-betareg(yield ~ batch + temp, data = GasolineYield)
summary(gy_logit)

gy_logit1 <- betareg(yield ~ batch + temp | batch + temp, data = GasolineYield)
summary(gy_logit1)

# Modelo probit

gy_probit <- betareg(yield ~ batch + temp, link='probit', data = GasolineYield)
summary(gy_probit)

gy_probit1 <- betareg(yield ~ batch + temp | batch + temp, link = 'probit', data
= GasolineYield)
summary(gy_probit1)

# Modelo cloglog

gy_cloglog <- betareg(yield ~ batch + temp, link='cloglog', data = GasolineYield)
summary(gy_cloglog)

gy_cloglog1 <- betareg(yield ~ batch + temp | batch + temp, link='cloglog', data
= GasolineYield)
summary(gy_cloglog1)

# Modelo loglog

gy_loglog <- betareg(yield ~ batch + temp, link='loglog', data = GasolineYield)
summary(gy_cloglog)
```

```
gy_loglog1 <- betareg(yield ~ batch + temp | batch + temp, link='loglog', data =
GasolineYield)
summary(gy_loglog1)

# Comparación de modelos por AIC

AIC(gy_logit, gy_logit1, gy_probit, gy_probit1, gy_cloglog, gy_cloglog1, gy_loglog,
gy_loglog1)

# Seleccion del mejor modelo dentro del seleccionado

gy_loglog2 <- betareg(yield ~ batch + temp | temp, data = GasolineYield)

library(lmtest)

lrtest(gy_loglog, gy_loglog2)
lrtest(gy_loglog2, gy_loglog1)

# Validacion de supuestos con el modelo logit

par(mfrow=c(3,2))
set.seed(123)
plot(gy_logit4, which = 1:4, type = 'pearson')
plot(gy_logit4, which = 5, type = 'deviance', sub.caption = " ")
plot(gy_logit4, which = 1, type = 'deviance', sub.caption = " ")
plot(gy_logit4, which = 6)

gy_logit4 <- update(gy_logit, subset = -4)
coef(gy_logit, model = 'precision')
coef(gy_logit4, model = 'precision')

AIC(gy_logit1, gy_logit4)


# Análisis a través de distancia


library(ecodist)

names(GasolineYield)
S <- dist(GasolineYield[,c("gravity","pressure","temp10",'temp')], method = 'eu-
clidean', diag = T, upper = T, p = 5)
Tc<- cmdscale(S, k = 5, eig = T)

yield1<- as.vector(GasolineYield$yield)
A<- as.matrix(-0.5*S)
Ide<- diag(1,nrow=length(yield1))
Jn<- matrix(rep(1, length(yield1)^2), nrow=length(yield1))/length(yield1)
H<- Ide-Jn
B<- H%*%A%*%H
Y<- eigen(B)

# Gráfico de valores propios y aporte
```

```
plot(Y$values)
sum(Y$values)
apor<- Y$values/sum(Y$values)*100; apor
plot(apor)

datos1< −data.frame(yield1,Y$vectors[,c(1:8)])
names(datos1)

datpd<- lm(yield1 ∼ X1+X2+X3+X4+X5+X6+X7+X8, data=datos1)
summary(datpd) AIC(datpd)

# Primer modelo con distancias

datpd_logit<- betareg(yield1 ∼ X1+X2+X3+X4+X5+X6+X7+X8, link = 'logit',
data = datos1)
summary(daty_logit)

datpd_logit1<- betareg(yield1 ∼ X1+X2 + X3+X4 + X5+X6 + X7+X8 | X1+X2
+ X3+X4 + X5+X6 + X7+X8, link = 'logit',data = datos1)
summary(datpd_logit1)

lrtest(datpd_logit,datpd_logit1)

AIC(datpd_logit,datpd_logit1)

datpd_logit1$fitted.values

# Segundo modelo con distancias

datpd_probit<- betareg(yield1 ∼ X1+X2+X3+X4+X5+X6+X7+X8, link='probit',
data=datos1)
summary(datpd_probit)

datpd_probit1<- betareg(yield1 ∼ X1+X2 + X3+X4 + X5+X6 + X7+X8 | X1+X2
+ X3+X4 + X5+X6 + X7+X8, link='probit', data=datos1)
summary(datpd_probit1)

lrtest(datpd_probit,datpd_probit1)

AIC(datpd_probit,datpd_probit1)

datpd_probit1$fitted.values

# Tercer modelo con distancias

datpd_cloglog<- betareg(yield1 ∼ X1+X2+X3+X4+X5+X6+X7+X8, link='cloglog',
data=datos1)
summary(datpd_cloglog)

datpd_cloglog1<- betareg(yield1 ∼ X1+X2 + X3+X4 + X5+X6 + X7+X8 | X1+X2
+ X3+X4 + X5+X6 + X7+X8, link='cloglog', data=datos1)
summary(datpd_cloglog1)

lrtest(datpd_cloglog, datpd_cloglog1)
```

```
AIC(datpd_cloglog, datpd_cloglog1)

datpd_cloglog1$fitted.values

# Cuarto modelo con distancias

datpd_loglog<- betareg(yield1 ~ X1+X2+X3+X4+X5+X6+X7+X8, link='loglog',
data=datos1)
summary(datpd_loglog)

datpd_loglog1<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 | X1+X2
+ X3+X4 + X5+X6 + X7+X8, link='loglog', data=datos1)

summary(datpd_loglog1)

lrtest(datpd_loglog,datpd_loglog1)

datpd_loglog1$fitted.values

AIC(datpd_logit, datpd_probit, datpd_cloglog, datpd_loglog)
AIC(datpd_logit1, datpd_probit1, datpd_cloglog1, datpd_loglog1)

# Cálculo de la disimilaridad

library(cluster)

Delta1<- daisy(GasolineYield[,-c(1)], metric ="gower")
class(Delta1)

mds1<- cmdscale(Delta1^(1/2), k = 31, eig = TRUE)
m<- sum(mds1$eig > 1,0e - 15)
mds1<- cmdscale(Delta1^(1/2), k = m, eig = TRUE)
names(mds1)

round(mds1$points[,1],4) # primera coordenada principal
plot(mds1$eig)

par(mfrow=c(1,2))
plot(mds1$points[,1], mds1$points[,2],
main=c('Representación en dimensión 2',"Batch"),
xlab='Coordenada principal 1', ylab='Coordenada principal 2',
col = c(1:10)[GasolineYield$batch], pch=c(1:10)[GasolineYield$batch], las=1)
abline(h=0)
abline(v=0)
legend(locator(1), c("batch 1", "batch 2", "batch 3", "batch 4", "batch 5", "batch
6", "batch 7", "batch 8", "batch 9", "batch 10"),
col=c(1:10), pch=c(1:10))

# Primer modelo con distancias

library(betareg)

ValoresPropios<- mds1$eig
X<- mds1$points
```

```
CorrCuadrado<- as.vector(cor(GasolineYield$yield,X)^2)
Porc.Inercia<-ValoresPropios/sum(ValoresPropios)
o<- data.frame(1:31, round(ValoresPropios,10), round(CorrCuadrado,10),
round(Porc.Inercia,10))
names(o)<- c('ID', 'ValoresProp', 'CorrCuad', 'Porc.Inercia')
fix(o)
plot(ValoresPropios, type='b')

sum(Porc.Inercia[1:12])

datos2<- data.frame(yield1, mds1$points[,1:12])
names(datos2)

# Modelo Logit

datpd_logit<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10
+ X11+X12, link='logit', data=datos2)
summary(datpd_logit)

datpd_logit1<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10
+ X11+X12 | X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10 + X11+X12,
link='logit', data=datos2)
summary(datpd_logit1)


lrtest(datpd_logit,datpd_logit1)

AIC(datpd_logit,datpd_logit1)

datpd_logit1$fitted.values

# Segundo modelo con distancias

datpd_probit<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10
+ X11+X12, link='probit', data=datos2)
summary(datpd_probit)

datpd_probit1<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10
+ X11+X12 | X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10 + X11+X12,
data=datos2, link='probit')
summary(datpd_probit1)

lrtest(datpd_probit, datpd_probit1)

AIC(datpd_probit, datpd_probit1)

datpd_probit1$fitted.values

par(mfrow=c(3,2))
set.seed(123)
plot(datpd_probit1, which = 1:4, type = "pearson")
plot(datpd_probit1, which = 5, type = "deviance", sub.caption = " ")
```

```
plot(datpd_probit1, which = 1, type = "deviance", sub.caption = " ")
plot(datpd_probit1, which = 6)

# Tercer modelo con distancias

datpd_cloglog<- betareg(GasolineYield$yield ~ mds1$points[,1:12], link='cloglog')
summary(datpd_cloglog)

datpd_cloglog1<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 +
X9+X10 + X11+X12 | X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10 +
X11+X12, link='cloglog', data=datos2)
summary(datpd_cloglog1)

lrtest(datpd_cloglog, datpd_cloglog1)

AIC(datpd_cloglog, datpd_cloglog1)

datpd_cloglog1$fitted.values

# Cuarto modelo con distancias

datpd_loglog<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10
+ X11+X12, link='loglog', data=datos2)
summary(datpd_loglog)

datpd_loglog5<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X8+X9 + X10
+ X12, link='loglog', data=datos2)
summary(datpd_loglog5)

lrtest(datpd_loglog, datpd_loglog5)

# Validación modelo 5

par(mfrow=c(3,2))
set.seed(123)
plot(datpd_loglog5, which = 1:4, type = "pearson")
plot(datpd_loglog5, which = 5, type = "deviance", sub.caption = " ")
plot(datpd_loglog5, which = 1, type = "deviance", sub.caption = " ")
plot(datpd_loglog5, which = 6)

# Modelo con ajuste finales

datpd_loglog1<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10
+ X11+X12 | X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10 + X11+X12,
link='loglog', link.phi='log', data=datos2)
summary(datpd_loglog1)

datpd_loglog2<- betareg(yield1 ~ X1+X2 + X3+X4 + X5+X6 + X7+X8 + X9+X10
+ X11+X12 | X1+X2 + X3+X4 + X5+X6 + X8+X10 + X11+X12, link='loglog',
link.phi='log',data=datos2)

lrtest(datpd_loglog, datpd_loglog1)
lrtest(datpd_loglog2, datpd_loglog1)
```

```
AIC(datpd_loglog,datpd_loglog1)

datpd_loglog1$fitted.values

# Comparación entre modelos y selección

AIC(datpd_logit, datpd_probit, datpd_cloglog, datpd_loglog)
AIC(datpd_logit1, datpd_probit1, datpd_cloglog1, datpd_loglog1)

# Modelo seleccionado log-log

datpd_loglog1$fitted.values

# Validación de supuestos modelo seleccionado

par(mfrow=c(3,2))
set.seed(123)
plot(datpd_loglog1, which = 1:4, type = "pearson")
plot(datpd_loglog1, which = 5, type = "deviance", sub.caption = " ")
plot(datpd_loglog1, which = 1, type = "deviance", sub.caption = " ")
plot(datpd_loglog1, which = 6)
```

Bibliografía

Arenas, C. & Cuadras, C. (2002), 'Recent Statistical Methods Based on Distances', *Constributions to Science, Institut d'Estudis Catalans Barcelona* **2**(2), 183–191.

Atkinson, A. (1985), *Plots, Tranformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford University Press, New York.

Bury, K. (1999), *Statistical Distributions in Engineering*, Cambridge University Press, New York.

Cook, R. (1986), 'Assessment of Local Influence (With Discussion)', *Journal of the Royal Statistical Society* **48**, 133–169.

Cook, R. D. (1977), 'Detection of Influencial Observations in Linear Regression', *Technometrics* **19**, 15–18.

Cribari-Neto, F. & Vasconcellos, K. (2002), 'Nearly Unbiased Maximun Likelihood Estimation for the Beta Distribution', *Journal of Statistical Computation and Simulation* **72**, 107–118.

Cuadras, C. (2007), *Métodos Multivariados Basados en Distancias*, Curso de doctorado, Universidad de Barcelona, Barcelona.

Cuadras, C. & Arenas, C. (1990), 'A Distance Based Regression Model for Prediction with Mixed Data', *Communications in Statistics A. Theory and Methods* **19**, 2261–2279.

Cuadras, C., Arenas, C. & Fortiana, J. (1996), 'Some Computational Aspects of a Distance-Based Model for Prediction', *Communications in Statistics. Simulation and Computation* **25**(3), 593–609.

Dobson, A. J. (2002), *An Introduction to Generalized Linear Models*, 2nd ed. Chapman Hall, Boca Raton, FL.

Espinheira, P. L., Ferrari, S. L. P. & Cribari-Neto, F. (2008), 'On Beta Regression Residuals', *Journal of Applied Statistics* **35**(4), 407–419.

Esteve, A., Boj, E. & Fortiana, J. (2010), 'Interaction Terms in Distance-Based Regression', *Communications in Statistics A. Theory and Methods* **38**(19), 3498–3509.

Ferrari, S. & Cribari-Neto, F. (2004), 'Beta Regression for Modelling Rates and Proportions', *Journal of Applied Statistics* **31**, 799–815.

Gower, J. (1968), 'Adding a Point to Vector Diagrams in Multivariate Analysis', *Biometrika* **55**, 582–585.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, John Wiley & Sons, vol.2, 2nd ed. New York.

Kieschnick, R. & McCullough, B. (2003), 'Regression analysis of variates observed on (0, 1): percentages, proportions, and fractions', *Statistics Model* **3**, 193–213.

Krysicki, W. (1999), 'On Some New Properties of the Beta Distribution', *Statistics and Probability Letters* **42**, 131–137.

Lee, Y. J. & Nelder, J. A. (2002), 'Analysis of Ulcer Data Using Hierarchical Generalized Linear Models', *Journal of Quality Technology* **21**, 191–202.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, London.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman Hall, London.

McCulloch, C. E. & Searle, S. R. (2001), *Generalized, Linear and Mixed Models*, John Wiley & Sons, New York.

Myers, R. H., Montgomery, D. C. & Vinning, G. G. (2002), *Generalized Linear Models. With Applications in Engineering and the Sciences*, John Wiley & Sons, New York.

Neter, J., Kutner, M. H., Nachtsheim, C. J. & Wasserman, W. (1996), 'A Tutorial on Generalized Linear Models', *Applied Linear Statistical Models* .

Ospina, R., Cribari-Neto, F. & Vasconcellos, K. L. P. (2006), 'Improved Point and Interval Estimation for a Beta Regression Model', *Computational Statistics & Data Analysis* **51**, 960–981.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992), *Numerical Recipes in C: The Art of Scientific Computing*, Second ed. Cambridge University Press, New York.

Simas, A., Barreto-Souza, W. & Rocha, A. (2010), 'Improved Estimators for a General Class of Beta Regression Models', *Computational Statistics & Data Analysis* **54**(2), 348–366.

Smith, D. M. & Ridout, M. S. (2003), 'Optimal Designs for Criteria Involving log(potency) in Comparative Binary Bioassays', *Journal Statistics Planning Inference* **113**, 617–632.

Smithson, M. & Verkuilen, J. (2006), 'A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Depedent Variables', *Psychological Methods* **11**(1), 54–71.

Vasconcellos, K. L. P. & Cribari-Neto, F. (2005), 'Improved Maximum Likelihood Estimation in a New Class of Beta Regression Models', *Brazilian Journal of Probability and Statistics* **19**, 13–31.

Wei, B. C., Hu, Y. Q. & Fung, W. K. (1998), 'Generalized Leverage and Its Applications', *Scandinavian Journal of Statistics* **25**, 25–37.